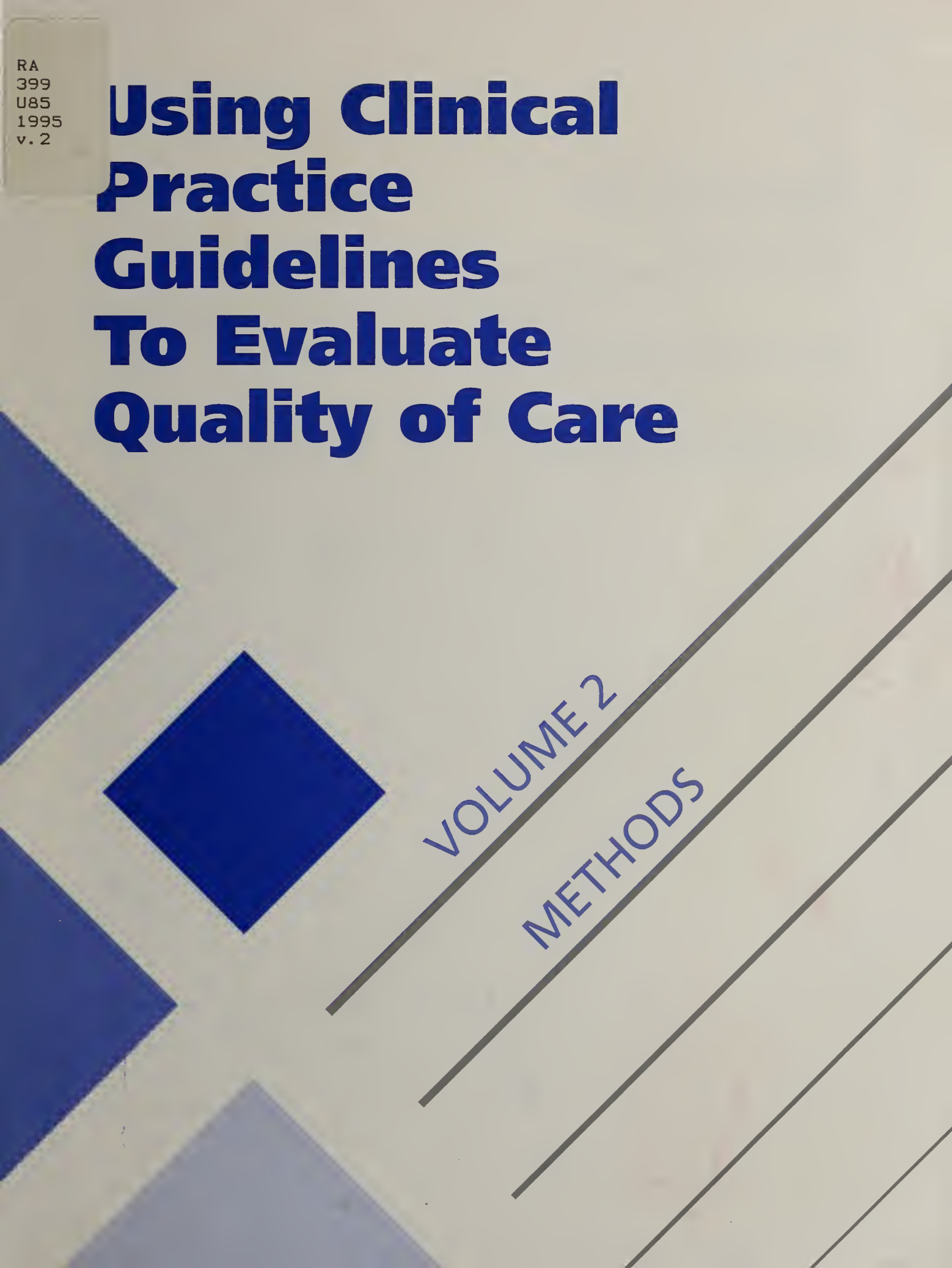
Using Clinical Practice Guidelines To Evaluate Quality of Care



Using Clinical Practice Guidelines To Evaluate Quality of Care

Volume 2: Methods

Workgroup Members

Stephen C. Schoenbaum, M.D., M.P.H., Co-chair
David N. Sundwall, M.D., Co-chair
David Bergman, M.D.
June M. Buckle, Sc.D., R.N.
Allan Chernov, M.D.
Janet George, R.N, B.S.N., M.A.
Clark Havighurst, J.D.
M. J. Jurkiewicz, M.D.
John T. Kelly, M.D., Ph.D.
Sandra Metzler, M.B.A., R.R.A.
Christine Miaskowski, Ph.D., R.N., F.A.A.N.
Sam J. W. Romeo, M.D., M.B.A.
Paul M. Schyve, M.D.
Bryan Simmons, M.D.
Patrice Spath, A.R.T., B.A.
Marcia Stevic, Ph.D., R.N.
Constance Winslow, M.D., M.B.A.
Steven Zatz, M.D.

U.S. Department of Health and Human Services
Public Health Service
Agency for Health Care Policy and Research

March 1995

AHCPR Pub. No. 95–0046

Foreword

The legislation that established the Agency for Health Care Policy and Research (AHCPR) in 1989 called for the agency to arrange for the development and periodic review and updating of:

1. Clinically relevant guidelines that may be used by physicians, educators, and health care practitioners to assist in determining how diseases, disorders, and other health conditions can most effectively be prevented, diagnosed, treated, and managed clinically; and

2. Standards of quality, performance measures, and medical review criteria through which health care providers and other appropriate entities may assess or review the provision of health care and assure the quality of such care.

As the process of developing clinical practice guidelines has evolved, the importance of the quality improvement tools that can be derived from guidelines has become clear, as has the need to publish and disseminate a series of reports on using guidelines to improve quality of care. In the applications illustrated in this series of reports, medical review criteria are explicit statements that clinicians, health care organizations, quality and utilization reviewers, payers, and regulators can use to determine how the process of care relates to guidelines. When derived systematically from credible guidelines, review criteria can be used to profile and assess quality of care. Criteria based on comprehensive guidelines can be used to examine how different aspects of the entire care process—diagnosis, treatment, and management—fit together, providing more useful information than isolated rates of incidence, utilization, or occurrence. Viewed in this way, medical review criteria are the building blocks for science-based performance measurement and standard setting in health care.

This two-volume report describes methodologies for translating AHCPR-supported clinical practice guidelines into review criteria and performance measures, and applications of those measures in quality of care standard-setting, assessment, and improvement. The studies included in this series illustrate the challenges of and innovative approaches to implementing clinical practice guidelines, measuring guideline-related performance, and assessing the effects of putting guidelines into practice. It is our hope that these reports will not only help health care professionals integrate guidelines into their own quality improvement programs, but will also stimulate discussion about concepts and methodology that will advance the field of quality measurement.

Clifton R. Gaus, Sc.D.
Administrator
Agency for Health Care Policy and Research

Preface

Change is occurring faster than ever before, or so it seems in health care in the United States. A restructuring is under way in the financing and delivery of health care, and at the national level new roles are being debated for the Government and the private sector in this huge enterprise that constitutes one-seventh of the economy.

While there are considerable differences of opinion among the many interested parties on just how this restructuring can best be accomplished, a clear consensus is emerging on some issues related to *quality* of care, such as the following:

- Quality of care is not a given.
- There should be full public accountability for outcomes of care.
- Accountability requires measuring the quality of care provided to populations of patients.
- Public policies should promote continuous improvement of quality of care.

This document is one effort by the Agency for Health Care Policy and Research (AHCPR) in fulfilling a legislative mandate to help achieve these new approaches to quality improvement, measurement, and monitoring. In short, this document was developed to help those interested in using clinical practice guidelines—developed by AHCPR and in the private sector—to evaluate quality of care. Although many view clinical practice guidelines primarily as tools in clinical quality improvement, interest has recently been growing in the application of clinical practice guidelines to broader purposes, such as payment and reimbursement policies and legal defense. Regardless of the purpose for which clinical practice guidelines are intended to be used, it is critical that everyone who uses them understands how to do so. For *any* judgment to be made about health care delivered according to recommendations based on clinical practice guidelines, such care must be evaluated with tools that have been rigorously developed on the basis of a clearly understood methodology.

Using Clinical Practice Guidelines To Evaluate Quality of Care describes how to translate clinical practice guidelines into evaluation tools and how to use these tools to determine whether the care provided conforms to the guidelines. The methodology is not simple, yet it is nonetheless conceptually straightforward. Good measurement of care—measurement that is credible to clinicians, managers, and the public—requires meticulous attention to rigorous technique. Therefore, readers are urged not to look for shortcuts in thinking through the issues and applying the techniques.

Using Clinical Practice Guidelines To Evaluate Quality of Care has been published in two volumes. Volume 1 consists of Chapters 1–3, which provide background and overview information about clinical guidelines, their potential usefulness in quality measurement, and issues that should be considered and

understood before proceeding to develop evaluation tools. Volume 2 (Chapters 4–6) contains detailed how-to instructions for developing guideline-derived evaluation tools. They describe the state of the art in the emerging science of quality measurement. While the recommendations may seem very detailed overall, the basic steps in the processes described can be applied to beginning efforts with limited resources as well as to mature quality measurement programs with ample resources.

The appendixes provide additional useful information related to complementary efforts in outcomes measurement, validity of performance measures, statistical ratio-based measurement, and algorithm development for review criteria.

Readers are encouraged to think through the issues in deriving measurements from guidelines and to apply the methods carefully to their own guideline-related measurement efforts. Only through additional efforts will the recommendations in this volume actually improve the methodology of care assessment and thereby improve care. The Nation is unquestionably poised on the threshold of this new era of accountability in health care.

Stephen C. Schoenbaum, M.D., M.P.H.
David N. Sundwall, M.D.
Workgroup Co-chairs

Abstract

As the United States undergoes a restructuring of the way health care is financed and delivered and debates the roles of the Government and the private sector in health care, there is growing interest in examining issues surrounding quality of care—how it is defined, how it is evaluated, and how delivery of care is related to patient outcomes.

A panel of experts with a broad range of expertise and experience in measuring quality of care was convened to develop a document that would provide a single methodology for using clinical practice guidelines to evaluate the quality of health care. Volumes 1 and 2 of this document describe how to translate clinical practice guidelines into evaluation tools and how to use these tools to determine whether the care provided conforms to the recommended guidelines. A fundamental assumption of this document is that such a methodology could have broad utility for a variety of potential users, including practitioners, providers, and professional organizations.

Workgroup Members

Stephen C. Schoenbaum, M.D., M.P.H.
Co-chair
Medical Director
Harvard Community Health Plan
New England Division
Providence, RI
Internist

David N. Sundwall, M.D.
Co-chair
President
American Clinical Laboratory Association
Washington, DC
Family Physician

David Bergman, M.D.
Department of Pediatrics
Lucile Salter Packard Children's Hospital
Palo Alto, CA
Pediatrician

June M. Buckle, Sc.D., R.N.
Director of Health Information Research and Development
GMIS
Malvern, PA
Nurse Researcher

Allan Chernov, M.D.
Vice President, Medical Services
Prudential Health Care System
Southwestern Group Operations
Sugar Land, TX
Internist

Janet George, R.N., B.S.N., M.A.
Director, Quality Management
Manor HealthCare Corporation
Silver Spring, MD
Nurse

Clark Havighurst, J.D. (resigned 1993)
Wm. Neal Reynolds Professor of Law
Duke University School of Law
Durham, NC
Attorney

M. J. Jurkiewicz, M.D.
Professor of Plastic Surgery, Emeritus
Emory University
Atlanta, GA
Surgeon

John T. Kelly, M.D., Ph.D.
Chief Medical Officer and Senior Vice President of Clinical Information Services
GMIS
Malvern, PA
Physician

Sandra Metzler, M.B.A., R.R.A.
Director, Resource and Outcomes Management
Columbia/HCA Healthcare Corporation
Nashville, TN
Health Information Specialist

Christine Miaskowski, Ph.D., R.N., F.A.A.N.
Associate Professor and Interim Chair
University of California San Francisco School of Nursing
Department of Physiologic Nursing
San Francisco, CA
Nurse Researcher

Sam J. W. Romeo, M.D., M.B.A.
Senior Associate Dean
University of Southern California
School of Medicine
Los Angeles, CA
Family Physician

Paul M. Schyve, M.D.
Senior Vice President
Joint Commission on Accreditation of Healthcare Organizations
Oakbrook Terrace, IL
Psychiatrist

Bryan Simmons, M.D.
Medical Director, Quality Management
Methodist Health Systems
Memphis, TN
Hospital Epidemiologist

Patrice Spath, A.R.T., B.A.
Partner, Brown-Spath & Associates
Forest Grove, OR
Health Information Specialist

Marcia Stevic, Ph.D., R.N.
Consultant
Sun Lakes, AZ
Health Services Researcher

Constance Winslow, M.D., M.B.A.
Consultant, Health Policy and Health Management
Berkeley, CA
Internist

Steven Zatz, M.D.
Senior Vice President
US Health Care
Blue Bell, PA
Internist

Methodology Consultants

R. Heather Palmer, M.B., B.Ch., S.M.
Director, Center for Quality of Care Research and Education
Harvard School of Public Health
Boston, MA

Naomi J. Banks, M.B.A., M.Ed.
Senior Research Analyst
Center for Quality of Care Research and Education
Harvard School of Public Health
Boston, MA

Ann G. Lawthers, Sc.D.
Lecturer
Harvard School of Public Health
Boston, MA

E. John Orav, Ph.D.
Associate Professor of Biostatistics
Department of Biostatistics
Harvard School of Public Health
Boston, MA

Technical Specialists

Alice G. Gosfield, J.D.
Alice G. Gosfield & Associates, P.C.
Philadelphia, PA
Attorney

Donald M. Nielsen, M.D.
Associate Medical Director and Quality Consultant
Permanente Medical Groups
Oakland, CA
Internist/Infectious Diseases

Writer/Editor

Suzanne Wymelenberg
Cambridge, MA

AHCPR Staff

Carole Hudgings, Ph.D., F.A.A.N.
Senior Health Policy Analyst
Office of the Forum for Quality and Effectiveness in Health Care

Jill Bernstein, Ph.D.
Chief, Program Planning, Research Development, and Evaluation Branch
Office of Program Development

Linda Demlo, Ph.D.
Director, Office of Program Development

Kathleen Hastings, R.N., J.D., M.P.H.
Director, Legal Medicine Program

Kathleen A. McCormick, R.N., Ph.D.
Senior Science Advisor
Office of Science and Data Development

Randie Siegel
Managing Editor
Center for Research Dissemination and Liaison

Federal Liaisons

Galen Barbour, M.D.
Associate Chief Medical Director for Quality Management
Department of Veterans Affairs
Washington, DC

Stephen Jencks, M.D.
Clinical Advisor
Health Standards and Quality Bureau
Health Care Financing Administration
Baltimore, MD

Michael J. Kussman, M.D.
Director, Medical Quality Assurance
OASD (HA) PAQA
Washington, DC

Acknowledgments

This document is the product of the entire workgroup, which brought to the table extensive experience with measurement of health care in a variety of settings and contexts. This experience proved invaluable in the delineation of important issues to be considered in developing and applying guideline-derived measures. The efforts of the following members of the workgroup who wrote drafts of specific issues are gratefully acknowledged: Dr. June Buckle, Dr. Allan Chernov, Mr. Clark Havighurst, Dr. John Kelly, Dr. Paul Schyve, Ms. Patrice Spath, and Dr. Marcia Stevic.

Dr. Heather Palmer and her colleagues, Ms. Naomi Banks, Dr. Ann Lawthers, and Dr. E. John Orav, played a crucial role in the work of the group and in the development of this document. Dr. Palmer and her colleagues had already worked out a basic method for developing review criteria and performance measures from guidelines and had tested it in the American Medical Review Research Center project. This project involved developing medical review criteria from three of the earliest AHCPR-supported clinical practice guidelines and the DEMPAQ project (the project to Develop and Evaluate Methods for Promoting Ambulatory Care Quality), supported by the Health Care Financing Administration. Her group, with input from all workgroup members and the assistance of the workgroup's medical writer, Ms. Suzanne Wymelenberg, wrote the chapters for Volume 2 on the translation methodology. Material for the appendixes was prepared by Dr. Palmer, Dr. Lawthers, and Ms. Banks (Appendix B, "Validity Review of Performance Measures"); Dr. Orav (Appendix C, "Statistical Issues for Rate-Based Measurement"); and Ms. Banks (Appendix D, "Constructing Algorithm Flowcharts for Performance Measure Evaluation").

Dr. Carole Hudgings from the staff of AHCPR served not only as the administrator who kept everyone on track but also as a key contributor to the substance of the work. Without her efforts and equanimity, the work could not have proceeded so smoothly, expeditiously, and productively.

Stephen C. Schoenbaum, M.D., M.P.H.
David N. Sundwall, M.D.
Workgroup Co-chairs

Contents

Executive Summary

To fulfill its legislative mandate to help achieve improvements in health care delivery, measurement, and monitoring, AHCPR convened a multidisciplinary expert panel to examine ways in which clinical practice guidelines can be used to evaluate quality of care.

Volume 1 (Chapters 1–3) presents a discussion of guideline-derived quality evaluation in which the basic tools for improving health care—medical review criteria, performance measures, and standards of quality—are briefly defined and described. Relevant issues that users should consider before developing and implementing these guideline-derived evaluation tools are discussed. The volume concludes with an appendix containing supplementary information on the use of outcomes as performance measures.

The heart of *Using Clinical Practice Guidelines To Evaluate Quality of Care* lies in Volume 2, in which Chapters 4–6 lay out an overview and explain step by step the methodology for translating clinical practice guidelines into evaluation tools. Medical review criteria, performance measures, and standards of quality are described in detail, and a discussion of the types of performance review, the skill requirements and costs, and the need for rigor in the methodology is presented.

Developing and implementing a guideline-derived performance measure is a process that involves three phases, each with multiple steps. These phases and steps are defined and described in detail, with multiple illustrative examples taken from AHCPR-supported clinical practice guidelines. First, the steps in the planning phase:

1. Clarify the purpose of the performance measurement.
2. Identify a relevant clinical practice guideline.
3. Identify populations covered by the guideline.
4. Identify guideline recommendations and draft the medical review criteria.

Second is the development phase:

5. Identify clinicians and sites of care.

6. Define case sample and case sampling period.

7. Identify data source.

8. Write medical review criteria, specifying acceptable alternatives and time window.

9. Specify data items and data rules.

10. Draft data collection forms and procedures.

11. Devise analysis procedures.

12. Pilot test and revise criteria, forms, and procedures.

And finally, the implementation phase:

13. Conduct review and assign criteria status.

14. Report review findings.

15. Interpret findings, apply standards of quality.

16. Investigate review findings.

17. Act on review findings.

18. Conduct review again to reevaluate performance.

Volume 2 includes a checklist (Chapter 6) that provides performance review committees with a means for checking the completeness of their work and documenting their decisions. Also included are appendixes with supplementary information on the development of a process to determine the validity of a performance measure, on selected statistical issues, and on methods for constructing algorithm flowcharts.

4. Overview of the Translation Methodology[1]

Concepts

This chapter presents an overview of developing evaluation tools—medical review criteria, performance measures, and standards of quality—based on clinical practice guidelines to evaluate quality of care. When these tools are based on clinical practice guidelines and the guidelines are based on scientific evidence analyzed according to an explicit methodology, a valid assessment of health care quality can result.

Since methods for health care quality and quality review derive from several clinical and administrative disciplines, there may be some confusion over technical terms; a term may have different meanings to persons with different backgrounds. Also, the same concept can be represented by several different words. To promote clear communication and efficient collaboration among health review professionals, and to make explicit the important differences in current interpretations, this chapter provides the standardized vocabulary adopted by the workgroup.

Medical Review Criteria

The workgroup adopted the following definition of medical review criteria: "Systematically developed statements that can be used to assess specific health care decisions, services, and outcomes" (see Volume 1, Table 2.1). Additional terms commonly used in discussing the development of medical review criteria are as follows, along with the definitions used by the workgroup:

- **Case mix.** Distribution of a group of patients into categories reflecting differences in patients' diagnoses/conditions.
- **Case severity.** A measure of intensity or gravity of a given condition or diagnosis for a patient.
- **Confidence interval.** An interval or range based on a random sample, for which there is a given probability (e.g., 95 percent) that the population mean is contained within that interval. For example, a study may

[1] *Authors: R. Heather Palmer, M.B., B.Ch., S.M.; and Naomi J. Banks, M.B.A., M.Ed.*

show that a drug lowers the average blood pressure for patients in the study by 4.8 mm Hg, with the 95-percent confidence interval between 2.5 and 7.3 mm Hg. The confidence interval is used in performance measurement to indicate whether an individual rate from a performance review is considered statistically similar to or different from the group average rate, or from a performance rate selected to represent an acceptable level of care.

- **Confidence limits.** The upper and lower boundaries of a confidence interval.
- **Explicit criteria.** Objective criteria specified in advance as a basis for making judgments of performance.
- **External review.** Review in which criteria and standards of judgment are developed or ratified with input from persons other than the clinician or clinician group that is being evaluated.
- **Implicit criteria.** Criteria formed by a respected clinician who uses clinical judgment in evaluating performance; these implicit criteria remain concealed in the mind of the reviewer.
- **Implicit review.** Review conducted using implicit criteria.
- **Internal review.** Review in which clinicians are involved in setting or adopting the criteria and standards by which they evaluate themselves.
- **Mean.** Arithmetic average of the values of a sample variable.
- **Peer review.** Review conducted by a peer (a similarly qualified clinician) or peers; historically, *peer review* has been done by *case-based implicit review*, and so the terms are sometimes used interchangeably.
- **Structured implicit review.** Implicit review conducted with instructions directing the reviewer to focus on certain types of data and answer certain questions in the review process.

Implicit Review

Quality assurance in health care began with the identification of single instances of poor quality through case-by-case review (for examples, see Richardson, 1972a, 1972b). A physician who was trusted to be an arbiter of clinical quality reviewed the data for a case and used his or her knowledge of appropriate clinical practice (or of an established guideline) to classify the care as acceptable or unacceptable. This is called implicit review, or "using implicit criteria," because the criteria on which the finding is based remain concealed in the mind of the reviewer. Implicit judgments have been shown to vary greatly (Brook and Appel, 1973; Goldman, 1992; Hayward, McMahon, and Bernard, 1993; Richardson, 1972a, 1972b; Rubin, Rogers, Kahn et al.,

1992; Sanazaro and Worth, 1985); the same unit of care may be evaluated differently by different reviewers or even by the same reviewer on different occasions. It has been common practice in quality-of-care research for the past 10 years to improve the reproducibility of implicit clinician reviews by carefully selecting and training such reviewers and by providing a "structured review instrument," which, based on specific data, contains directions for clinician judgments about specific issues (Brook and Appel, 1973; Rubenstein, Kahn, Reinisch et al., 1990; Rubin, Rogers, Kahn et al., 1992; Sanazaro and Worth, 1985; see also Appendix B.)

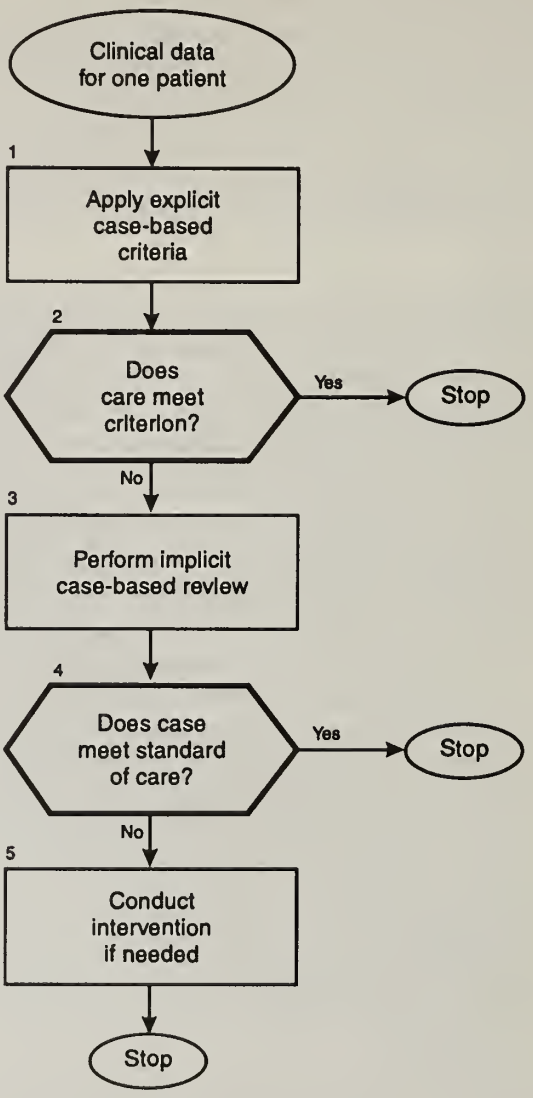
Explicit Review

In the 1970s, efforts were made to improve the reproducibility of case reviews and to reduce costs by conserving clinicians' time in the following way: Clinicians wrote down criteria for judging quality (explicit criteria); these criteria were applied by nonclinician reviewers; clinicians then implicitly reviewed only the cases that failed to meet the written criteria. (For an example, see Richardson, 1972a, 1972b). Figure 4.1 shows this sequence of review for a single case (the term *standard of care* in box number 4 is defined in the "Standards of Quality" section in this chapter). The wording in the legislation that established AHCPR and in the Institute of Medicine's (IOM's) definition of medical review criteria applies to *explicit* medical review criteria (IOM, 1990).

Before the advent of formal clinical practice guidelines, individuals or organizations wishing to conduct performance reviews implied or constructed a guideline or protocol when they defined the clinical content of explicit medical review criteria. Now that published clinical practice guidelines are available, explicit review criteria can be written to measure conformance to those guideline recommendations. How closely medical review criteria adhere to the clinical practice guideline from which they are derived is a measure of the validity of the criteria. The guideline itself is validated against patient outcomes by the nature of the development process, which is based on scientific evidence and expert opinion.

The attributes of medical review criteria suggested by the IOM Committee on Clinical Practice Guidelines as desirable (Table 4.1 and IOM, 1992) can be summarized as follows. Sensitivity and specificity are especially important characteristics for review criteria. The workgroup modified the explanations of these terms given by the IOM to illustrate use of review criteria to detect conformance to a guideline. A review criterion is sensitive if it detects conformance in a high proportion of cases that truly do conform to the guideline. In other words, a criterion that is sensitive produces a high rate of true positive identifications of conformance to guideline and a low rate of false negative identifications of nonconformance. A criterion is specific if it detects nonconformance in a high proportion of cases that truly do not conform to guideline. That is, a criterion is specific if it produces a low rate of false positive identifications of conformance to guideline and a high rate of true negative identifications of nonconformance.

Figure 4.1. Conducting case-based review



The IOM (1992) also advises that criteria should allow for patient preference in choosing between alternative tests and treatments. Further, it recommends that complex review criteria be translated into formats that are understandable to nonphysician practitioners, reviewers, patients, and other consumers of health care. Medical review criteria should not be used to place excessive burdens on clinicians or patients. The data required to evaluate criteria conformance should be easily obtained. Review criteria should be presented with explicit instructions for application and scoring, and these instructions should be formulated so that they can readily be transformed into computer-based review systems. When medical review criteria are used in a way that negatively affects providers (e.g., for payment decisions or sanctions), there should be provision for appeal of adverse review decisions.

Table 4.1. Desirable attributes of medical review criteria

Attribute	Explanation
Sensitivity[1]	A review criterion intended to test for conformance to a guideline is sensitive if the criterion classifies as conforming to guideline a high proportion of cases that truly do conform to the guideline.
Specificity[1]	A review criterion intended to test for conformance to a guideline is specific if the criterion classifies as nonconforming a high proportion of cases that truly do not conform to the guideline.
Patient responsiveness	Review criteria specifically identify a role for patient preferences or the process for using them allows for some consideration of patient preferences.
Readability	Review criteria are presented in language and formats that can be read and understood by nonphysician reviewers, practitioners, and patient/consumers.
Minimum obtrusiveness	Review criteria and the process for applying them minimize inappropriate direct interaction with and burdens on the treating practitioner or patient.
Feasibility	The information required for review can be obtained easily from direct communication with providers, patients, records, and other sources, and the decision criteria are easy to apply. Review criteria are accompanied by explicit instructions for application and scoring.
Computer compatibility	Review criteria are straightforward enough that they can be transformed readily into the computer-based protocols and similar formats that can make the review process more efficient for all involved parties.
Appeals criteria	Criteria provide explicit guidance about the considerations to be taken into account when adverse review decisions are appealed by professionals or patients.

[1]Explanations of these terms were modified by the workgroup to illustrate use of review criteria to test for conformance to a guideline.

SOURCE: Adapted from the Institute of Medicine, Committee on Clinical Practice Guidelines. Guidelines for clinical practice: from development to use. Field MJ, Lohr KN, editors. Washington, DC: National Academy Press, 1992.

Relationship of Guidelines to Medical Review Criteria

It is important to distinguish between clinical practice guidelines and the medical review criteria that are derived from them. Clinical practice guidelines and medical review criteria differ in four important ways (Table 4.2).

Purpose. Guidelines have a prospective focus: they are designed to assist clinicians and patients in making decisions about health care to be given. Medical review criteria, however, assess care decisions that have already been made. Medical review criteria may be used retrospectively, when the care decided on has already occurred, or prospectively, when the decision has been made but not yet implemented, as in preprocedure and surgical second opinion reviews.

Table 4.2. Differences between guidelines and medical review criteria

	Clinical practice guidelines	Medical review criteria
Purpose	Guide care to be given	Evaluate decisions already made
Data	Data obtained as required	Use data documenting care given
Care sequence covered	Cover all pathways	Cover main pathways
Role of clinical judgments	Complement clinical judgment	Allow for clinical judgment

SOURCE: Adapted from S. Greenfield. Measuring the quality of office practice. In: Goldfield N, Nash DB, editors. Providing quality care. The challenge to clinicians. Philadelphia: American College of Physicians, 1989.

Data. Guidelines recommend patient data that should be obtained as they are needed to make clinical decisions; medical review criteria are applied to data that were collected in the course of delivering care. During performance review, data not already documented are not usually obtained after the fact. Exceptions would be some forms of preprocedure review and second surgical opinions. For example, reviewers using online computerized programs may ask clinicians for additional information needed to determine whether the stated medical review criteria are met. When such prospective application of medical review criteria elicits further relevant clinical facts, the quality of the decision for or against a procedure may be improved.

Care sequence covered. A sequence of management recommendations in a guideline may follow a set of branching pathways corresponding to variation in the health states of the patients concerned. Since reviewing all pathways, including those that affect few patients, is expensive, medical review criteria are generally limited to main pathways such as those dealing with high-risk, high-volume, or problem-prone areas known or suspected to need improvement.

Role of clinical judgment. The IOM recommends that clinical practice guidelines include specifically known or generally expected exceptions to the recommendations. This is the desirable guideline attribute "clinical flexibility," as specified by the IOM (IOM, 1990; see Volume 1, Table 3.2). However, it is not practical for a guideline to include all possible exceptions to the recommendations, especially for the rarest combinations of patient circumstances. For some patients, therefore, clinicians may justifiably decide that the clinical practice guideline does not apply. Because medical review criteria are written to determine whether or not the guideline was followed, allowance must be made for situations in which the recommendations of the guideline do not apply or need not be followed. Such common expressions of clinical judgment, called "acceptable alternatives" in this text, may be known elsewhere as "exceptions."

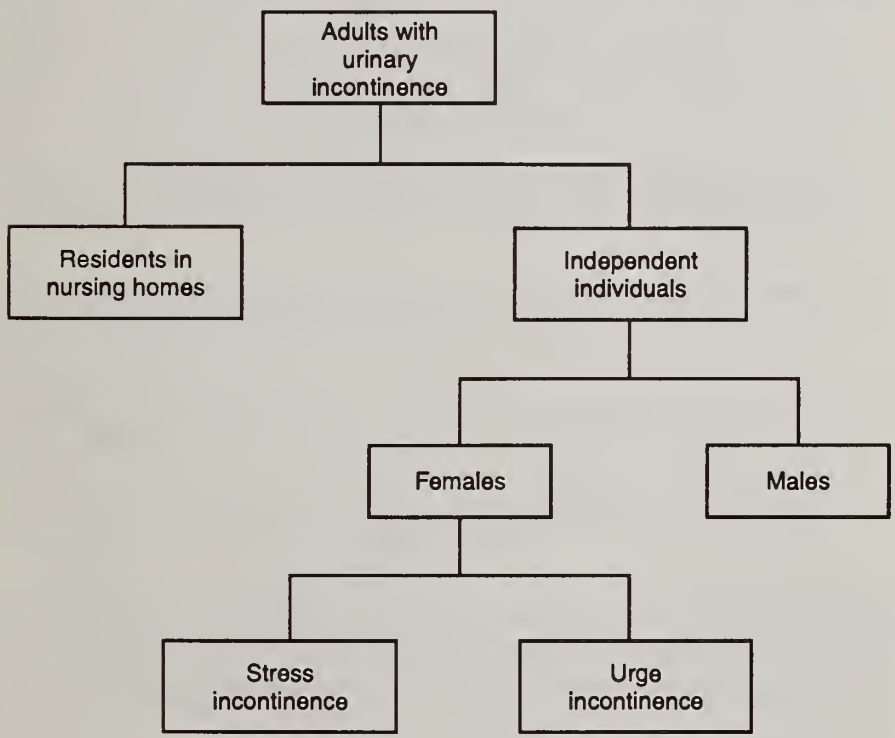
Criteria Sets Derived From Clinical Practice Guidelines

Since practice guidelines address appropriate treatment of a condition or disease for many types of patients in a variety of circumstances, many diverging

pathways or sequences of care are described. For example, the AHCPR-supported clinical practice guideline *Urinary Incontinence in Adults* (Diokno, McCormick, Colling et al., 1992) includes branches that discuss treatment of patients in nursing homes and in the community. Further branching takes place depending on gender and type of incontinence (Figure 4.2). At each branching point, the guideline recommendations become more specific for a more narrowly defined and homogeneous group of patients.

Each review criterion describes a discrete management decision or action, administrative structure, or health outcome. Effort is wasted applying criteria derived from every recommendation in a guideline to a heterogeneous group of patients, because each criterion would apply only to a small number of patients. Instead, the natural branching points of the guideline can be used to define criteria sets. Each set includes medical review criteria based on that portion of a guideline that applies to a narrowly defined group of patients. For the guideline on urinary incontinence, for example, several criteria sets could be developed, depending on nursing home residency versus independence, male versus female, and (for females) stress incontinence versus urge incontinence.

Figure 4.2. Urinary incontinence guideline: branching to form criteria sets



Performance Measures

The term *performance measures*, as it is used in this document and by the workgroup, is defined as follows: " Methods or instruments to estimate or monitor the extent to which the actions of a health care practitioner or provider

conform to the clinical practice guideline" (Volume 1, Table 2.1). Other terminology used frequently in the discussion of performance measures is as follows:

- **Performance indicators.** Quantitative measures used to measure and improve performance and quality (JCAHO, in press). *Rate-based* indicators are similar to the *performance measures* defined in Volume 1, Table 2.1; they produce rates for comparing the performance of organizational providers of care. *Sentinel event* indicators identify undesired events such as death; a single instance of a sentinel event triggers a quality review (JCAHO, 1990, p. 11; JCAHO, 1991, p. 43). In referring to rate-based indicators, this text uses the term *performance measures,* as does the legislation that established the AHCPR.

- **Performance rates.** Measurements produced by using a performance measure, providing a quantitative evaluation of quality of patient care (see **Rate**, below).

- **Population-based.** The word *population* in this instance is used in the epidemiological sense—a defined group of individuals sampled for study. The term *population-based* is sometimes used to mean *rate-based*. However, a population-based measure generally is a performance rate for a group of patients in a geographic area or in a particular health plan enrollment. When used in this sense, *population-based* applies to all patients rather than to only those who use services.

- **Profiles.** Sets of performance rates aggregated by clinician, clinician group, or organization to monitor some aspect of health care delivery.

- **Rate.** A quantitative measure, usually expressed as a percentage, of the occurrence of an event of interest within a specified time interval. Rates are derived by creating a fraction in which the numerator is the number of patients experiencing an event of interest and the denominator is the population of patients at risk for the occurrence of that event. A rate may also be constructed by counting events rather than patients in the numerator and denominator—that is, when the event could occur more than once for a given patient.

- **Variation.** For performance rates, *variation* refers to differences between the performance rate of one clinician or group of clinicians or organizations and the performance rates of comparable others.

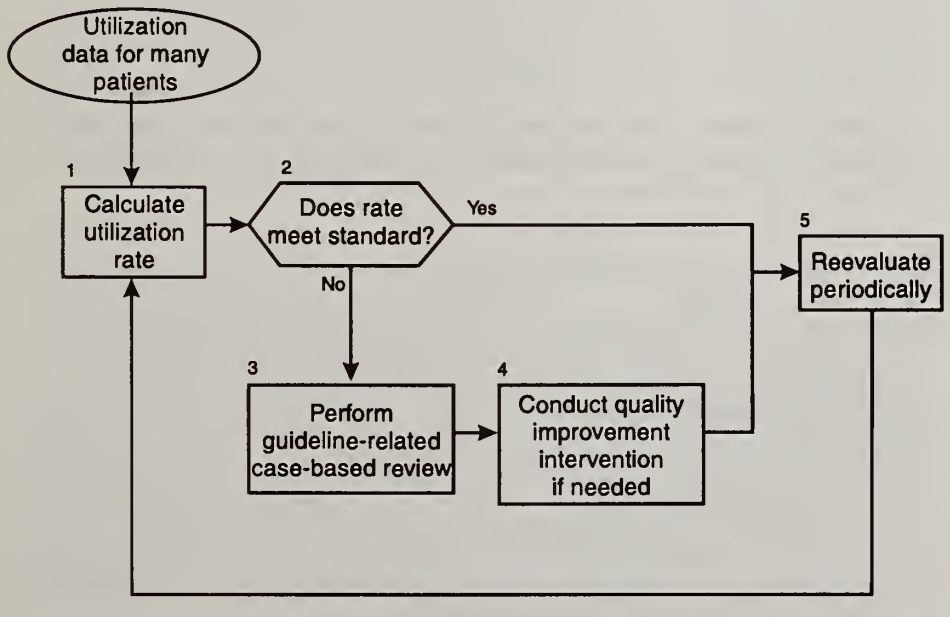
Reviewing Utilization Rates

Performance measures meld together into one tool two different approaches for evaluating quality during the past decade, namely, reviews of individual cases and analyses of variation in aggregate utilization rates. Traditional, implicit, case-based reviews focus on individual instances of care and commonly use patient records as a data source. Variation in the rates of surgical proce-

dures as observed in claims and other administrative data files has emerged as an alternative means of examining quality of care (Wennberg, 1986). For example, physicians who have higher rates of utilization for a procedure than their peers may be overusing the procedure. To the extent that such comparisons reveal true differences in appropriate use of the procedures, the rates can be considered approximate indicators of quality of care. Utilization rates are primarily useful as indicators of overuse and underuse; they are relatively coarse measures of quality, since they do not determine the status of each case in terms of its conformance to a guideline. However, as discussed below, differences in utilization rates can be investigated by applying explicit guideline-derived medical review criteria to individual cases (see, for example, Chassin, Kosecoff, Park et al., 1987).

Utilization rates can be used for performance review as follows: The performance evaluation process may be viewed as a cycle (Figure 4.3) that may not necessarily start with a clinical practice guideline. For example, a performance evaluation can begin with the construction of the cesarean section (C-section) rate for organization A. This rate then can be compared to a standard of quality derived from the C-section rate of other, similar organizations. If organization A is found to have a C-section rate higher than that of its peer organizations, the hospital quality improvement committee may investigate why this is so. The committee may choose a clinical practice guideline for C-sections and develop from it medical review criteria and a performance measure. By reviewing whether cases in organization A receive care in conformance to the guideline, the committee determines how much of the difference in C-section rates reflects inappropriate care. This determination also may suggest remedial efforts to decrease the rate of unwarranted C-sections.

Figure 4.3. Review of utilization rates



Remeasurement of the hospital's C-section rates can then answer the questions "Have we improved?" and "Now how do we compare with others?"

While this approach is a legitimate one and fairly common, this document concerns quality performance measurements that *begin* with intent to assess conformance to a clinical practice guideline. These performance measurements apply explicit medical review criteria derived from a practice guideline to individual cases, then aggregate the case review results to construct performance rates. This approach is described below.

Measuring Conformance to Practice Guidelines

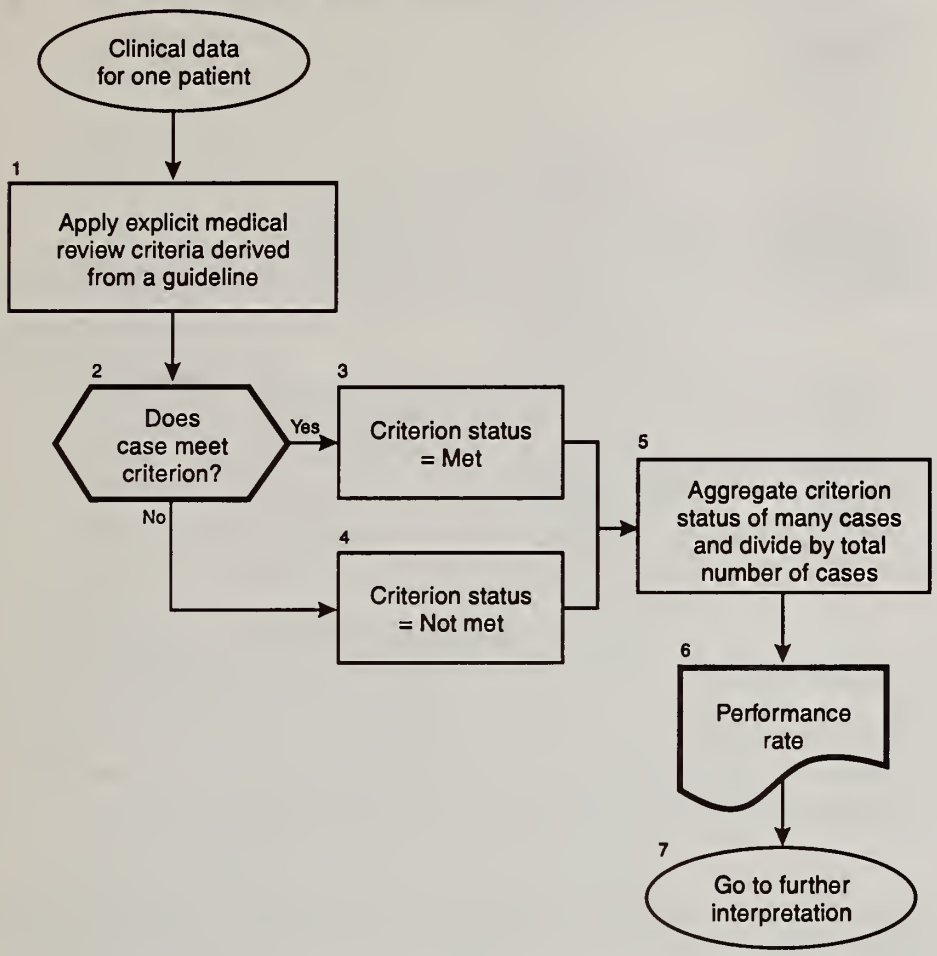
The mandate for clinical practice guidelines, medical review criteria, performance measures, and standards of quality is directed toward ensuring effective care for patients while reducing inappropriate variation in use of services. Good clinical practice guidelines provide clear recommendations for what should be done for particular patients, and related medical review criteria are used to determine whether particular patients received the recommended care. The methodology described here melds the two traditional types of performance review that were described above, namely, review based on services delivered to individual cases and review based on rates of delivery of services to populations of patients. The two methods are combined into a single tool—the performance measure—for measuring guideline conformance.

In using a performance measure, results from explicit criterion review of many cases are aggregated to form an average rate of performance. Once the results of individual case reviews are collected in a data base, average performance rates can be computed for individual clinicians or further aggregated to reflect the performance of health care organizations or of all the health care organizations in a region. Creating rates of conformance to a guideline based on data from many cases is useful, because any single case may be exceptional and therefore may not accurately reflect the average performance of a clinician, health care organization, or group of organizations. Performance rates constructed by applying guideline-derived medical review criteria to many cases create a more accurate picture of the average performance of a clinician, organization, or group of organizations. Figure 4.4 illustrates how performance rates are derived from aggregated case assessments.

Components of a Performance Measure

Observing a single event permits only an isolated quality judgment. To measure the quality of care patients receive, it is necessary to devise an instrument—the performance measure—which, like a ruler to measure length, will reveal an amount, or rate, of "quality" of care given to a population of patients (Palmer, 1991). Since quality cannot be observed directly, an event that is evidence of quality of care is selected. Such events or processes are specified in the medical review criteria derived from a practice guideline; medical review criteria are thus a key component of a performance measure for evaluating conformance to the guideline recommendations.

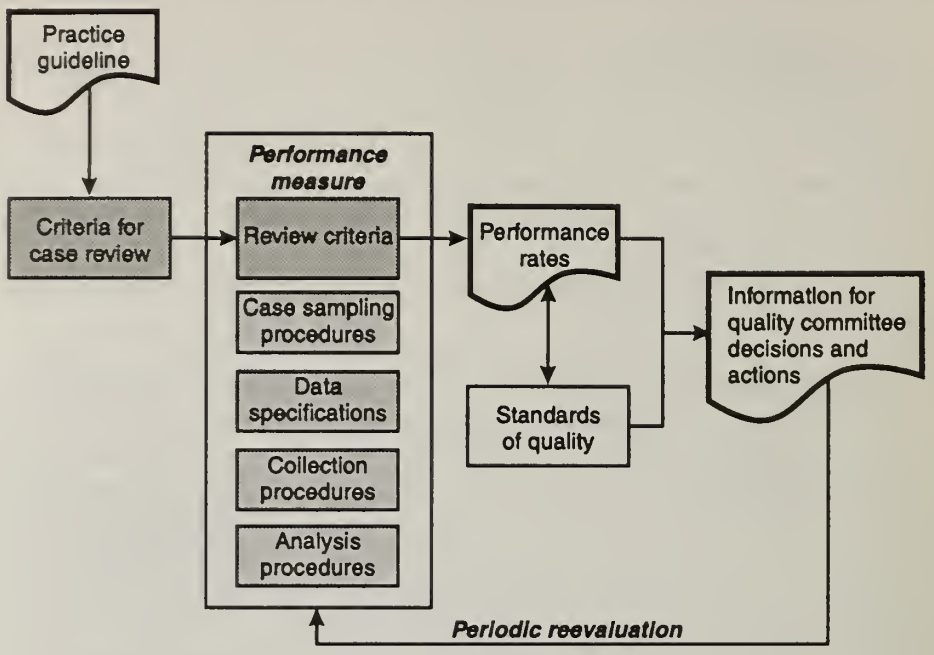
Figure 4.4. Applying explicit medical review criteria to cases to construct a performance rate



In addition to medical review criteria, the four other components of a performance measure are (1) the case sampling procedures, (2) the specifications for the data to be collected, (3) the data collection procedures, and (4) the analysis procedures. Figure 4.5 shows the relationship of these components to one another and to practice guidelines and standards of quality.

Correct specifications for all components of a performance measure are essential to ensure reliability and validity of the measure. For precision and reproducibility in measurement, an instrument must be constructed so that each unit of measurement has the same value, in the way that each inch on a ruler is equal to all other inches. In measuring clinical care, the unit of measurement is a case, and all cases assume equal value in the performance measurement. To ensure that each case is equivalent in the important characteristics relevant to the practice guideline, the performance measure specifies (1) the type of patient for whom the criteria will apply (the case sample); (2) the exact data items that will be examined to determine guideline conformance; (3) the precise procedures for collecting data; and (4) the procedure for analyzing the

Figure 4.5. Relationship of clinical practice guidelines to review criteria, performance measures, and standards of quality



data. If all cases are not equivalent with respect to these characteristics, the performance measure is not precise, and measurement error (see the "Measurement Error" section) will undermine confidence in use of the instrument.

A performance measure is a tool for producing a quantitative measurement of quality of care expressed as a performance rate. However, simply knowing a rate does not reveal whether or not it is acceptable. A judgment of the acceptability of a performance rate must be made in relation to the purpose for which it will be used. Thus, the final stage in measuring health care quality is applying a standard of quality that embodies some concept of the acceptability of a particular performance rate for a specific purpose.

Standards of Quality

The IOM definition of standards of quality adopted by the workgroup is as follows: "Authoritative statements of: (1) minimum levels of acceptable performance or results, (2) excellent levels of performance or results, or (3) the range of acceptable performance or results" (Volume 1, Table 2.1). Other terms often used in a discussion of such standards are as follows:

- **Benchmark.** A level of care set as a goal to be attained. Internal benchmarks are derived from similar processes or services within an organization; competitive benchmarks are comparisons with the best external competitors in the field; and generic benchmarks are drawn from the best performance of similar processes in other industries.

- **Comparative standard.** A standard derived from a comparison with other performance rates constructed by using exactly the same performance measure, such as the prior performance of a clinician or provider, the observation of the performance of others, or the statistical analysis of group rates.

- **Prescriptive standard.** A statement of what should be achieved rather than a statement of what has been achieved.

- **Standard for accreditation.** A statement of expectation set by competent authority concerning a degree or level of requirement, excellence, or attainment in quality or performance (JCAHO, in press).

- **Standard of care (legal usage).** In malpractice case court proceedings there is an attempt to determine whether a patient suffered harm due to negligent violation of a standard of care. The standard of care for the case is elaborated by the questioning of expert witnesses who have studied the facts of the case that are before the court and have relevant knowledge of comparable behavior.

- **Standard of care (regulatory usage).** Standards for facilities are commonly expressed in terms of a minimal level of policy, equipment, and capacity necessary to achieve licensure or certification.

- **Threshold.** A preestablished level for care. If a desired attribute of care falls below this level or an undesired attribute of care rises above this level, further evaluation or action is triggered.

Relationship of a Standard of Quality to a Review Criterion for a Single Case

Standards of quality are used with performance measures to decide whether intervention concerning quality of care is necessary. To avoid confusion, the meaning of the term *standard* must be clarified when assessing a single case or many cases. For a single case, case-based standards are used to determine whether intervention is needed. For a performance measure derived from review of many cases, rate-based standards are used to direct interventions.

When case-based review is done, a single case is classified by the review as receiving or not receiving good-quality care. The simplest form of standard setting is an implicit case-based review by a peer reviewer. In addition to use for quality review by hospitals and professional review organizations (PROs), peer review is conducted by physician reviewers when a preprocedure review decision is appealed. When a powerful, negative intervention hangs on the decision in a single case, the workgroup believes that due process should include additional review and the opportunity for appeal before a final determination is made that the case-based standard was not met.

Explicit criteria derived from a clinical practice guideline imply a prescriptive standard of care based on the guideline recommendations. However, guidelines and their related review criteria cannot apply to every case, because the explicit criteria may not cover every possible combination of patient circumstances for every case that is reviewed. The correct clinical course for a given case may be an exception to the recommendation given in the guideline. Such a case would then be wrongly classified by guideline-related medical review criteria as receiving inappropriate care. In such an instance, if it is essential to determine whether mitigating circumstances account for the lack of conformance to the criteria, an additional review is conducted to determine whether the case truly received inappropriate care. This sequence was illustrated in Figure 4.1.

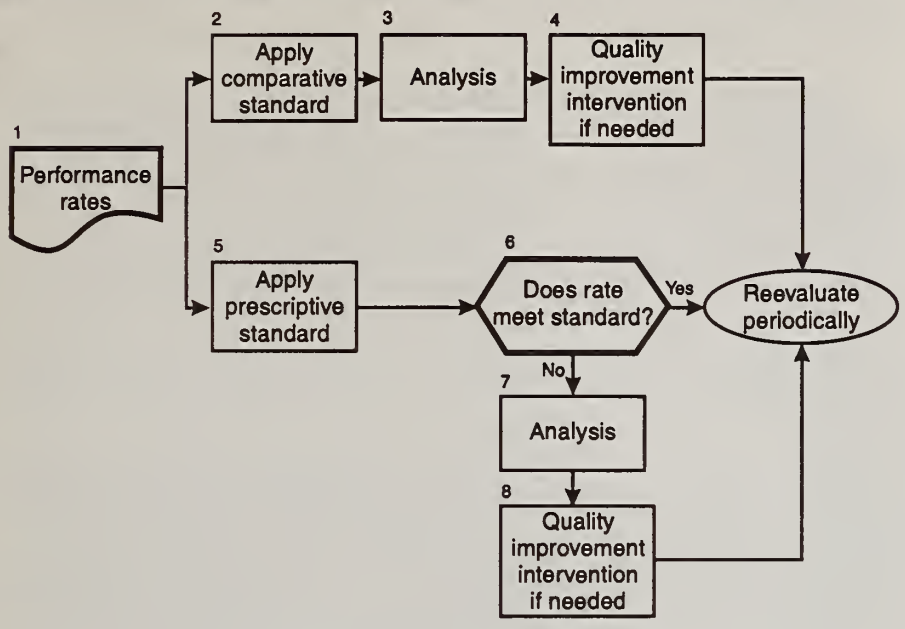
When the status of many cases reviewed by explicit criteria is expressed as a performance rate, errors of judgment about particular cases are less crucial than in an individual case review. If a rate is based on the review of many cases, confidence in the meaning of the rate is not compromised if a few cases are erroneously classified. If the number of wrongly classified cases is few, there is no great effect on the rate. However, to have confidence in a rate as a valid means to measure performance, it is important to know the usual percentage of cases that may be erroneously classified by that rate. This percentage is an estimate of the measurement error associated with this measurement instrument. For further discussion of this point, see the section "Need for Rigor" and Appendix B.

Standards of Quality for Performance Measurement

Standards of quality are applied to performance rates (Figure 4.6.). In this context, a standard is the means of deciding whether the rate of conformance to the guideline requires further investigation in order to determine whether to take action to improve health care quality, and what action to take. As indicated by the definition of standards of quality, different standards can be applied to a given performance rate, depending on the purpose of the review. For instance, a minimum standard for a desired attribute of care may be established, using either the prescriptive or the comparative approach; if the performance rate falls below this threshold, further investigation is triggered and may lead to action to improve quality. Similarly, for an undesired attribute of care, either the prescriptive or the comparative approach may be used, and the threshold defines an upper limit; exceeding this limit triggers further investigation and, possibly, action (James, 1989).

For example, suppose that a hospital committee finds that its rate of inappropriate use of coronary angiography is 5 percent. Does this finding justify immediately diverting resources to a series of interventions to improve appropriateness? In a 1981 research study, rates of inappropriate use of this procedure in several regions of the United States averaged 17 percent (Chassin, Kosecoff, Park et al., 1987), and in a 1990 study the rate of inappropriate use of this procedure in New York State was 4 percent (Bernstein, Hilborne, Leape

Figure 4.6. Applying standards of quality to a performance rate



et al., 1993). Knowing of these studies, and also knowing that quality improvement resources are urgently needed in other areas, the committee may decide in advance that a rate of inappropriate use of coronary angiography of 10 percent or less would be considered acceptable for its purpose. Thus, no interventions are launched once the 5-percent rate of inappropriate use is determined, so that scarce resources may be directed to quality improvements in areas the committee deems more urgent. In this example, a rate of 10 percent or less is the standard of quality.

A comparative standard of quality can be used to identify the best conformance to guidelines achieved by any clinician, group of clinicians, or group of organizations. This type of standard is a benchmark or goal. It is used to spur others to reach that level of performance. Comparisons of guideline conformance over time within an organization, or at one point in time between organizations, can be used in a continuous quality improvement program. The initial measurements of performance are compared with the benchmark to determine the priorities for designing system improvements. Followup measurements can indicate whether system changes have indeed improved performance.

The comparative approach can also be used to set upper and lower bounds for guideline conformance. Comparisons of guideline conformance are first made between groups of clinicians. Typical questions might be "How well do internists versus family physicians treat patients with heart failure?" or "How do pediatricians in Virginia compare with the pediatricians in Georgia on management of acute asthmatic attacks?" Statistical analysis is then applied to the range of performance rates to determine which groups of clinicians differ from others by more than chance variation. (For technical details, see Appendix C.)

Comparisons of guideline conformance among providers or over time may also reveal trends or patterns of performance that should trigger further investigation of opportunities for improvement.

The use of such comparative standards will become more important in the competitive managed care environment. Health maintenance organizations (HMOs), individual practice associations, and other health care plans vying for employers and members can use performance measures and standards of quality to distinguish their services and also to direct quality improvement. Purchasers of care can use such comparisons to purchase "value"—that is, to obtain the best care possible for the dollars spent.

Use of Standards of Quality in Relation to Scientific Evidence for a Guideline

Performance measures may reasonably be derived from clinical practice guideline recommendations for which the scientific evidence is not strong, particularly if the recommendation is strongly made. There are many examples of clinical issues on which scientific evidence is weak but the clinical indications to intervene are strong. For instance, the AHCPR-supported guideline for unstable angina diagnosis and management (Braunwald, Mark, Jones et al., 1994) recommends use of morphine sulfate for patients whose symptoms are not relieved after three serial sublingual nitroglycerin tablets or whose symptoms recur with adequate anti-ischemic therapy; this recommendation is made despite the lack of trials testing use of morphine for these specific purposes.

When a guideline recommendation is strong despite weak scientific evidence, review criteria are still written and assessed as "met" (i.e., "This case meets the review criterion") or "not met" ("This case does not meet the review criterion"). The resulting performance rate can be used to guide further analyses and quality improvement efforts. The standard of quality may be set below 100 percent if the guideline recommendation is based on weak or conflicting scientific evidence, since expecting 100-percent conformance would seem unreasonable.

In other instances, a guideline may contain a recommendation that is not strong despite the existence of strong scientific evidence. Such instances are when (1) the evidence shows potential benefits for patients to be small in relation to the potential risks or discomfort, or (2) the magnitude of the potential benefit may be very small. Nevertheless, in some situations constructing a measure of conformance to such a recommendation is useful. However, since many patients may refuse such a procedure, it would be reasonable to apply a standard of quality of far less than 100 percent in deciding whether to take further action on the basis of actual performance results.

Overview of Methods

Types of Performance Reviews

There are methodological and logistical differences between reviews conducted within a single organization, such as a hospital or HMO, and those conducted for clinicians or organizations defined by a geographic area (e.g., State, PRO region) or membership group (e.g., members of the American Academy of Family Physicians; staff clinicians in a large, multicenter HMO or preferred provider organization). Today, there is a rapid trend toward multiorganization performance measures. The measures are compiled by using computerized review systems and are deployed across many organizations. For instance, the Department of Defense, the Department of Veterans Affairs, the Health Care Financing Administration (HCFA), and the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) already use such systems or are building them (Barbour, 1994; Halperin, in press; JCAHO, 1990, 1991; Mayer, Clinton, and Newhall, 1988; Wilensky, 1991). An increasing number of private companies market similar systems. An organization can use such a computerized system for performance measurement in fulfilling its responsibility for quality improvement. For example, in 1994 the JCAHO introduced an Indicator Measurement System in which, on a voluntary basis, any JCAHO-accredited hospital can collect data using the indicators in the system, and send the data to the Joint Commission. The JCAHO will send back to each participating hospital a comparison of its performance with that of similar hospitals. The hospital will then be able to use its own performance rates, compared with those of its peers, to guide its quality improvement activities.

Despite this trend, many clinicians and organizations continue to conduct their own internal performance measurements, using their own resources. Therefore, in discussing the methods for developing medical review criteria, performance measures, and standards of quality from practice guidelines, it is important to remember that the basic methods are the same no matter who develops the performance measures.

Steps in Conducting Guideline-Based Performance Measurement

No matter which type of review is done, 18 steps or tasks can be identified in the process of developing a measure of guideline conformance (Table 4.3). These steps are explained in detail in Chapter 5. In the planning phase there should be a focus on the purpose of the review. The appropriate guideline to satisfy that purpose and the population to whom the guideline applies are identified (steps 1–3) (American Medical Association, 1992). Next, the specific guideline recommendations or logically complete portion of the guideline relevant to the purpose of review are selected, and the medical review criteria are drafted in outline form (step 4).

The methods and conduct of the review are planned in the development phase. The clinicians and sites of care are specified along with the

Table 4.3. Steps in developing and implementing a guideline-derived performance measure

Planning phase	1. Clarify the purpose of the performance measurement
	2. Identify a relevant clinical practice guideline
	3. Identify populations covered by the guideline
	4. Identify guideline recommendations and draft the medical review criteria
Development phase	5. Identify clinicians and sites of care
	6. Define case sample and case sampling period
	7. Identify data source
	8. Write medical review criteria, specifying acceptable alternatives and time window
	9. Specify data items and data rules
	10. Draft data collection forms and procedures
	11. Devise analysis procedures
	12. Pilot test and revise criteria, forms, and procedures
Implementation phase	13. Conduct review and assign criteria status
	14. Report review findings
	15. Interpret findings, apply standards of quality
	16. Investigate review findings
	17. Act on review findings
	18. Conduct review again to reevaluate performance

characteristics of the sample of patients whose care is to be reviewed (steps 5–6). The patient data sources available at these sites of care are identified so that specific medical review criteria may be considered (step 7). Criteria are developed in specific detail, reflecting decisions made in steps 5–7. These criteria are grouped into a criteria set encompassing the individual patient care decisions and the actions and outcomes from the guideline that the performance measure assesses (step 8).

Medical review criteria, in order to be truly guideline derived, must not exclude guideline recommendations that are related to one another. For example, a guideline may recommend against surgery for benign prostatic hyperplasia for patients with mild symptoms but may recommend a careful followup of the patient (sometimes called "watchful waiting"). The medical review criteria for this guideline should state both that surgery should not be done and that careful followup should be done. Acceptable alternative conditions under which specific criteria need not be met also must be defined (step 8). Data needed for the review are identified; data abstraction forms and the rules for data collection are written; analysis procedures are devised; and then all forms and procedures are tested and, if necessary, revised (steps 9–12).

After the performance measure has been tested and finalized, the implementation phase involves actually doing the review (step 13); reporting and evaluating the resulting performance rates against an appropriate standard (steps 14–15); doing the followup evaluation and taking any necessary corrective action (steps 16–17); and, ultimately, repeating the review to evaluate whether performance has improved (step 18).

Note that steps 6–11 in Table 4.3 specify the five components of a performance measure shown in Figure 4.5: the medical review criteria, the specifications for the sample of patients to be reviewed, the specifications for data to be collected, the procedures for that data collection, and the analysis procedures. Note also that step 15 in Table 4.3 is to interpret the review findings; doing so requires consideration of standards of quality for this performance measure. The application of standards of quality to performance measurements is also illustrated in Figure 4.5.

As decisions are made at each stage of developing a performance measure, the range of choices available at subsequent steps diminishes. That is, prior choices constrain further ones. For example, in examining conformance to the AHCPR-supported guideline on urinary incontinence in adults (Diokno, McCormick, Colling et al., 1992), the decision to focus on recommendations related to nursing home care (step 4) eliminated outpatient clinicians and provider sites from the review (step 5). If a further decision is made that the guideline recommendations for review will concern the appropriateness of surgery that was performed (step 4), then the patient sample (step 6) will be limited to patients who have had surgery. Since this type of surgery generally requires hospital admission, the data sources (step 7) will include not only nursing home patient records but also hospital patient records.

The focus of this document is on the methodological steps—steps 3 to 14—that must be understood in the planning, development, and implementation of medical review criteria and performance measures. In addition, attention is given to understanding the use and interpretation of standards of quality in relation to performance rates (step 15). The 18 steps in Table 4.3 can be identified whether the review is an implicit review of patient records to investigate the utilization of coronary angioplasty or a multi-State comparison of adherence by urologists, internists, and family physicians to the AHCPR-supported guideline *Benign Prostatic Hyperplasia: Diagnosis and Treatment* (McConnell, Barry, Bruskewitz et al., 1994). Although the steps may not always seem to be discrete or to occur in exactly the order described here, they can be identified for any type of review, as discussed below.

Contrast Between Case-Based Review and Performance Measurement

Case-based reviews, by definition, end with assessing the care given to individual cases. For instance, the review may be triggered by the interest of a group of clinicians in their own conformance to a guideline or by the results

of guideline-related performance measurement. The purpose of case-based review is to answer a question about the quality of care for a specific case or for a group of cases that may have one or more characteristics in common, such as the same provider, diagnosis, or iatrogenic outcome. In case-based review, by definition, no attempt is made to construct a rate by aggregating data from a group of cases. The qualitative interpretation of the judgments made for each case reviewed provides guidance for quality improvement. This type of case-based review has traditionally formed the basis of the monitoring conducted by the PROs. Now, however, PRO reviews take the form of analysis of rates, profiles, and patterns of care (Jencks and Wilensky, 1992).

Performance measurement may be conducted by an organization to apply guideline-derived explicit review criteria to a concern raised by utilization review or as an ongoing program of quality monitoring. In this instance, the performance rates produced by the measurement apply only to the single organization itself. If this measurement suggests a quality problem, the problem often can be confirmed or denied through further review and investigation.

Multiorganizational performance measurement, derived from large numbers of similar cases, allows comparisons over time and among health care delivery sites. The profiles and patterns of care revealed by performance measurements permit organizations to prioritize efforts to improve overall performance. Increasingly, computerized systems and services for comparative performance measurements are available to individual organizations that do not themselves have the resources for planning and implementing such measurements.

Skill Requirements

In general, the steps shown in Table 4.3 may involve four types of skills: clinician and nonclinician management skills, clinical expertise, technical expertise in performance measurement, and health care information management expertise. Participation in performance review requires expertise in the methods and issues discussed in this document. Any participants who do not already possess this expertise will benefit from specific training. The four types of skills are discussed in the remainder of this section.

Individuals with *clinician or nonclinician management* expertise may identify a need for the review, receive the results of the review with recommendations for action, and decide what action, if any, to implement. In the hospital or HMO setting, nonclinician managers may serve as members of a quality improvement committee.

In the case of external reviews initiated by government policymakers (HCFA, State departments of health), nongovernment payers or purchasers (managed care corporations, employers), and specialty societies, performance measurements are calculated and presented to clinician and nonclinician managers of a health care organization for their assessment and action decisions.

Clinical expertise is needed to ensure that an appropriate clinical practice guideline is selected and that the medical review criteria and performance measures that are produced relate to the recommendations in that guideline. Clinical expertise is also used in setting standards of quality for evaluating the performance rates that are obtained. At the organizational level, clinical experts may be members of a quality improvement committee or an ad hoc task force. If the standing committee lacks representation from disciplines relevant to the guideline, professionals in those disciplines should be recruited for the development of the performance measure.

To develop performance measures applicable to multiorganizational review, a committee or expert panel is convened. The panel members are not necessarily officials of the body that convened the panel, but they are academic and community-based practitioners known for their practical skill, judgment, clinical experience, and excellence in knowledge. Experience in evaluation of quality of care is also essential. It is important that the panel members be widely recognized and respected by the group of clinicians whose performance is to be evaluated with the performance measure. Relevant credentials include appropriate professional training and an active role in professional societies. Panel members should include representation from all health care disciplines responsible for the guideline-recommended care. For example, an expert panel for acute pain management might include a neurologist, a surgeon, an anesthesiologist, a critical care or primary care nurse, an internist, a physical therapist or rehabilitation specialist, and a clinical pharmacologist. If the results will be applied to a State, a region, or the Nation, panel members should be drawn from different geographic areas, from urban and rural locales, and from different types and sizes of organizations. However, panel size is best kept small so that discussion and the group work do not become unwieldy.

Technical expertise is required to ensure the methodological integrity of the performance measure and a sound approach to its implementation. At the organizational level, this expertise may be provided by a health information management professional or a quality assurance/quality improvement professional. Large-scale studies require staff skilled in data management and data analysis to construct the performance rates.

Multiorganizational performance measurement presents methodological issues that require more highly trained personnel, usually scientists with master's degrees or doctorates. Producing measurements that have the required reliability and validity to convince large numbers of clinicians and policymakers of their usefulness requires using more rigorous methods, as discussed below. Technical consultants on the expert panels must be experienced in study design, sampling issues, data base design, data analysis, and computerization of analysis algorithms. For these projects, individuals with management skills are also needed to organize and direct the review at the many sites or among the many participating clinicians.

For performance measures within an organization, *health care information management specialists* (formerly known as medical record specialists) or nurse reviewers can examine records and develop explicit instructions for finding and coding specified data items. When a large staff is responsible for abstracting data from patient records from many different organizations, a senior staff member is required to train and supervise the data abstractors. Increasingly in the future, data downloaded from health care information systems may be used for measuring guideline conformance. Data system experts are essential for this type of performance measurement program.

Costs of Review

Personnel costs are a large component of any project involving patient record review. Although the clinician reviewers who conduct implicit case-based reviews may command high hourly fees, this method is generally applied to many fewer cases than are examined in performance measurement. If highly paid clinicians were to review each of the large number of cases required for performance measurement, the review would become prohibitively expensive. It is common practice, therefore, for health care information specialists or nurse reviewers to abstract the required data, using explicit medical review criteria. This approach reduces the cost per case to reasonable levels.

For small organizations unable to marshal the resources to develop their own systems for performance measurement, vendors of review services now provide the full range of necessary skills and tools. By marketing these services to many customers, vendors offer sophisticated review services at lower cost to the small organization than the organization would expend by conducting its own review. For multiorganizational reviews, the cost of each review is significantly reduced by the economies of scale obtained from using automated technology when a performance measure must undergo a rigorous development and testing process.

In addition to the salaries of the information specialists, the following expense items also may be necessary for multiorganization performance measurement:

- Expert committee or panel: travel, meals, lodging, meeting facilities, communications, and compensation.
- Computerization of the performance measurement: programming, data entry, equipment, program updates, and maintenance.
- Administration: reproducing patient records for reviews not done on site, travel to participating organizations, record room fees for pulling and refiling records, and charges for computer searches of claims and other administrative data files.
- Reporting performance measurements: producing profiles, printing and distributing reports, and communicating and explaining rates.

In planning performance measurements, the volume of cases has an important bearing on the methods employed. The type of abstracting form or data analysis method that suits a single organization may not be suitable for a multiorganization effort. In Chapter 5, considerations of cost and efficiency are included in the discussions of methods for each step in conducting performance measurements.

Including Rigor in the Methodology

Need for rigor. Implicit review, as conducted in past decades, is not methodologically rigorous. It has been shown that implicit review judgments vary among reviewers (Brook and Appel, 1973; Goldman, 1992; Hayward, McMahon, and Bernard, 1993; Richardson, 1972a, 1972b; Rubin, Rogers, Kahn et al., 1992; Sanazaro and Worth, 1985). However, when implicit review is used in an organization, the reputation of the peer reviewer among colleagues lends authority to the findings. Although single implicit judgments may be useful for quick assessments of quality among colleagues, the workgroup believes that penalties should not be imposed on the basis of the findings of a single reviewer without a "due process" that involves opportunities for appeal and rebuttal of the findings.

When developing a performance measure, issues of sensitivity, specificity, validity, and reliability should be considered. This point is especially important when the review covers many organizations, geographic locations, or professional groups. Even within a single organization, the rigor of the review methodology is important, and rigor may be difficult to achieve. Thus, the reliability of the data abstraction process should be assessed regularly by reabstracting a sample of data for comparison with the original abstraction and investigating causes for any disagreement between the two versions.

When the development and use of a performance measure is said to be rigorous, the implication is that its review criteria comply with certain attributes recommended by the IOM (see Table 4.1; the "Explicit Review" section of this chapter; and IOM, 1990, 1992). That is:

- The measure actually measures conformance to the clinical practice guideline on which it is based (validity).
- The measure can differentiate between care that adheres to clinical practice guidelines and care that does not (sensitivity and specificity).
- The instrument gives the same results for identical cases when the measurement is repeated by the same abstractor or another abstractor, when a measurement is repeated sometime after the initial assessment, and whether the cases come from the same organization or from across the country (reliability).

These attributes are essential for measures that compare performance of different organizations or at different points in time. It is unwise for an

organization to undertake significant interventions designed to change performance unless the measures have high validity and reliability. If repeated performance measurement is used to monitor and evaluate the effectiveness of interventions, it is important to assess whether the reliability and validity of the measures have changed during the period of study.

Methodological rigor is attained by identifying measurement error and reducing it insofar as the funds available for review permit. Testing the performance measure is discussed below and in Chapter 5.

Measurement error. Measurement provides an estimate of an attribute (e.g., quality), and estimates always have some error. When variations in a series of measurements are found, the differences usually are interpreted to mean that real differences in quality exist. If the measure were error free, that interpretation would be true. However, causes other than true variations in the quality of care may account for the variation. These variations constitute measurement error. Reducing error in the estimates of performance made by a performance measure is an important function of a quality review committee or expert panel. Decisions made in drafting the data collection forms and procedures (step 10, Table 4.3), the data analysis procedures (step 11), and in revising the measure after pilot testing (step 12) are directed toward reducing measurement error.

Error may be either random or systematic. If a measure is reliable, repeated measurements of the same thing yield similar results (i.e., the rate of random error is low). A reliable measure, however, may have limited validity because of a high rate of systematic error. Systematic error exists when a measure relates to an attribute other than the one intended. Reliability and validity of measures can be illustrated by the analogy of trying to tune to a particular radio station. Receiving static noise is equivalent to a high rate of random error (i.e., the signal is not received reliably). Tuning to the wrong station results in receiving a clear signal (low random error) but is equivalent to an invalid signal (a signal other than the one intended). When the instrument measures what it is intended to measure, both random and systematic error are low and the measurement is said to be valid (equivalent to tuning in to the radio station that was originally sought).

Measurement errors are greater when the measurement is based on small numbers of cases. It is therefore more difficult and more expensive to achieve valid measurements at the level of the individual clinician, for whom the available sample of cases may be small, than at the group level.

Among the major sources of measurement error are inaccuracies in the data used for performance measurement. Data errors can occur in recording, coding, collecting, or transmitting data about patient care events and in interpreting such events during further coding and analysis. Error from these sources can be reduced by scrupulous attention to data quality control, includ-

ing precise definition of the data sample, data specifications, data collection rules, and procedures for data analysis.

Measurement error may also arise because it is impossible to write medical review criteria in such detail that they apply to every possible combination of patient circumstances. If the patient care under review can be influenced by known sources of variation other than the intended quality attribute, the review criteria being used can be revised to exclude such error sources. For example, measurement error can be minimized by writing into the medical review criteria the obvious exceptions to a guideline recommendation—known as "acceptable alternatives"—such as the patient's refusing tests or treatment. It is also possible in statistical analyses of performance measurements to adjust for variations that are related to factors other than quality differences. For example, performance measures can be adjusted for patient characteristics that do not indicate a different course of care but that may affect the patient's willingness to accept the recommended care or the difficulty for the clinician to provide such care (McNeil, Pedersen, and Gatsonis, 1992; Orav, Louis, Palmer et al., 1991).

It is particularly important to consider statistical adjustments of performance rates in conducting profile and pattern analysis. In an HMO, hospital, or nursing home, many professionals share in the care of one patient, and their efforts are influenced by the system of care in which they work. Measuring the effect of any single contributor to care requires adjusting for the effects of all others.

Easiest to measure is the performance of a whole system, which includes the contributions of all the individuals in it. The main source of measurement error then is variation contributed by the patients whose care is studied and by the circumstances of the time and place of the care that was given. When the performance of a group of clinicians is to be measured, a source of error is introduced by including events that are controlled even in part by the system, the patients, or the time and place of the care given (American Hospital Association, 1991; Quality Measurement and Management Project, 1991, pp. 8–9). For example, if the event being measured is clinician performance in recommending mammograms, the actual completion of mammograms may be used to determine performance rates. However, the number of mammograms completed can also be affected by the system's refusal to pay for such screening, the patient's refusal to have a mammogram, or the travel time required to reach the radiology department. If these reasons do not equally affect all clinicians, performance measures that do not adjust rates when mammograms are not done for these reasons cannot be used as fair measures of a clinician's performance.

References

American Hospital Association. Practice pattern analysis: a tool for continuous improvement of patient care quality. Chicago: American Hospital Association; 1991.

American Medical Association. Using practice parameters in quality assessment, quality assurance, and quality improvement programs. Chicago: American Medical Association; 1992.

Barbour GL. Development of a quality improvement checklist for the Department of Veterans Affairs. The Joint Commission Journal on Quality Improvement 1994;20(3):127–39.

Bernstein SJ, Hilborne LH, Leape LL, et al. The appropriateness of use of coronary angiography in New York State. JAMA 1993;269:766–9.

Braunwald E, Mark DB, Jones RH, et al. Unstable angina: diagnosis and management. Clinical Practice Guideline. AHCPR Pub. No. 94–0602. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services; March 1994.

Brook RH, Appel FA. Quality-of-care assessment: choosing a method for peer review. N Engl J Med 1973;288:1323–9.

Chassin MK, Kosecoff J, Park RE, et al. Does inappropriate use explain geographic variations in the use of health care services? JAMA 1987;258:2533–7.

Diokno A, McCormick K, Colling J, et al. Urinary incontinence in adults. Clinical Practice Guideline. AHCPR Pub. No. 92–0038. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services; March 1992.

Goldman RL. The reliability of peer assessments of quality of care. JAMA 1992;267:958–60.

Greenfield S. Measuring the quality of office practice. In: Goldfield N, Nash DB, editors. Providing quality care. The challenge to clinicians. Philadelphia: American College of Physicians; 1989.

Halperin J. The measurement of quality of care in the Veterans Health Administration (VHA). Medical Care Supplement. In press.

Hayward RA, McMahon LF, Bernard AM. Evaluating the care of general medicine inpatients: how good is implicit review? Ann Intern Med 1993;118:551–7.

Institute of Medicine (IOM), Committee on Clinical Practice Guidelines. Clinical practice guidelines: directions for a new program. Field MJ, Lohr KN, editors. Washington, DC: National Academy Press; 1990.

Institute of Medicine (IOM), Committee on Clinical Practice Guidelines. Guidelines for clinical practice: from development to use. Field MJ, Lohr KN, editors. Washington, DC: National Academy Press; 1992.

James BC. Quality management for health care delivery. Quality Measurement and Management Project. Chicago: The Hospital Research and Educational Trust; 1989.

Jencks SF, Wilensky GR. The health care quality improvement initiative: a new approach to quality assurance in Medicare. JAMA 1992;268(7):900–3.

Joint Commission on Accreditation of Healthcare Organizations (JCAHO). An introduction to quality improvement in health care. Oakbrook Terrace, IL: JCAHO; 1991.

Joint Commission on Accreditation of Healthcare Organizations (JCAHO). Primer on indicator development and application: measuring quality in health care. Oakbrook Terrace, IL: JCAHO; 1990.

Joint Commission on Accreditation of Healthcare Organizations (JCAHO). Lexikon: a dictionary of health care terms, organizations, and acronyms for the era of reform. Oakbrook Terrace, IL: JCAHO; in press.

Mayer W, Clinton JJ, Newhall D. A first report of the Department of Defense external civilian peer review of medical care. JAMA 1988;260(18):2690–716.

McConnell JD, Barry MJ, Bruskewitz RC, et al. Benign prostatic hyperplasia: diagnosis and treatment. Clinical Practice Guideline No. 8. AHCPR Pub. No. 94–0582. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services; February 1994.

McNeil BJ, Pedersen SH, Gatsonis C. Current issues in profiling quality of care. Inquiry 1992;29:298–307.

Orav EJ, Louis TA, Palmer RH, et al. Variance components and their implications for statistical information in medical data. Stat Med 1991;10:599–616.

Palmer RH. Part I. Considerations in defining quality of health care. In: Palmer RH, Donabedian A, Povar GJ, editors. Striving for quality in health care: an inquiry into policy and practice. Ann Arbor, MI: Health Administration Press; 1991. p. 1–54.

Quality Measurement and Management Project. Hospital quality-related data: recommendations for appropriate data requests, analysis, and utilization. Chicago: The Hospital Research and Educational Trust; 1991.

Richardson FM. Methodological development of a system of medical audit. Med Care 1972a;10:451–62.

Richardson FM. Peer review of medical care. Med Care 1972b;10:29–39.

Rubenstein LV, Kahn KL, Reinisch EJ, et al. Changes in quality of care for five diseases measured by implicit review, 1981 to 1986. JAMA 1990;264:1974–9.

Rubin HR, Rogers WH, Kahn KL, et al. Watching the doctor-watchers: how well do peer review organization methods detect hospital care quality problems? JAMA 1992;267(17):2349–54.

Sanazaro PJ, Worth RM. Measuring clinical performance of individual internists in office and hospital practice. Med Care 1985;23:1097–114.

Wennberg JE. Which rate is right? N Engl J Med 1986;314:310–1.

Wilensky GR. From the Health Care Financing Administration: Medicare's PROs change their focus, broaden their mission. JAMA 1991;266:2810.

5. Designing and Testing Medical Review Criteria and Performance Measures[1]

Introduction

When reviewers did implicit case-based review in recent decades, the reviewer's own knowledge base provided the implicit review and standard of care for a case. The method was convenient and satisfied a need until scientifically based practice guidelines and methods for more objective explicit review were developed. This chapter recounts how practice guidelines change implicit case-based review and describes the design and testing of guideline-based performance measurements.

Implicit Review of Guideline Conformance

There are many variations in methods for conducting implicit case-based reviews; for instance, a "structured implicit review" method was briefly described in Chapter 4. (Published examples of that method can be found in Rubenstein, Kahn, Reinisch et al., 1990, and Rubin, Rogers, Kahn et al., 1992.) An alternative "structured implicit review," used to evaluate validity of performance measures, is described in Appendix B. One common approach is described in this section.

When implicit case-based review is used to investigate a utilization rate or quality issue, the clinician-reviewer engages mentally in a process that incorporates the steps listed in Table 4.3.

The reviewer identifies a clinical practice guideline that is relevant to the patient or patients concerned (steps 2 and 3). If this clinical practice guideline is already integrated into the reviewer's knowledge base, it may not be formally consulted during the implicit review (step 4). The clinicians, staff, and sites of care to be reviewed, as well as the patients whose health records (the usual data source for implicit review) will be examined, are determined according to the purpose of the review. For example, a review of high rates of cesarean section begins with the patients cared for by the obstetric department. Similarly, a review of a series of adverse drug reactions might involve examining the records of patients hospitalized on the units where those reactions occurred (steps 5–7).

[1]*Authors: R. Heather Palmer, M.B., B.Ch., S.M.; and Naomi J. Banks, M.B.A., M.Ed.*

The clinician does not identify in advance the data items to be sought in a patient record or a procedure for applying the criteria to patient data (steps 8–12). Yet, because the reviewer has internalized the guidelines, the reviewer will note items documented in the care of the patient that indicate conformance or nonconformance with the guideline. For instance, in a review of adverse drug reactions, the reviewer should seek evidence that the medications administered were indicated, that there were no contraindications, and that there was adequate monitoring for significant drug side effects. These criteria, although unwritten, would be derived from the reviewers' internalization of relevant clinical practice guidelines.

When conducting a review (step 13), the clinician mentally combines the purpose of the review, the facts of the case, and familiarity with the pertinent guideline from which he or she has implicitly formulated criteria for acceptable performance. After making a judgment about whether these implicit criteria have been adhered to, the reviewer makes a report to the proper committee regarding the quality of the care for each case reviewed (step 14). At that point, further judgment is made to interpret the review findings according to the prevailing standards of quality in the organization (step 15). Questions such as these may be asked: Were too many cesarean sections performed without appropriate indications? Does it appear that medications are being inappropriately prescribed or insufficiently monitored? The committee thus implicitly applies a standard based on priorities within the organization to determine whether further action should be taken (steps 16 and 17).

Method for Measuring Guideline Conformance

This section describes measurement of guideline conformance. There are various methods for conducting such measurements (American Hospital Association, 1991; Spath, 1990, 1992a, 1992b). A generic methodology approved by the workgroup is described here. This chapter uses the example of developing a performance measure either by a single health care organization (hospital, long-term care facility, HMO, group practice) or by a group of many organizations, for the purpose of quality improvement.

A single organization generally would put together a quality assurance/quality improvement (QA/QI) committee (or performance review committee) to follow the steps described in Table 4.3 for planning, developing, and implementing a performance measure. The steps used by a single organization are illustrated in this chapter by a sequence of examples to show how they would apply to different types of care. The examples are derived from the AHCPR-supported clinical practice guideline number 4, *Cataract in Adults: Management of Functional Impairment* (O'Day, Adams, Cassem et al., 1993) and are shown in a series of shaded boxes.

The methods for developing performance measures that encompass many health care organizations are similar to those for single health care organiza-

tions, but because of the large number of cases involved and the variability in practice between different organizational and geographic settings, additional issues arise. In this chapter, the text describing the additional procedures that are necessary for dealing with many organizations is indented and italicized.

There are many different ways to develop and implement measurement of guideline conformance (Longo and Bohr, 1991; Miller and Knapp, 1979; Palmer, Louis, Hsu et al., 1985). One set of options is described here. The example relates to an expert panel as it follows the steps outlined in Table 4.3.

A multidisciplinary expert committee or panel is used to derive medical review criteria, performance measures, and standards of quality from a practice guideline. As with the AHCPR-supported guideline panels themselves, the validity of the product depends on the expertise and practical experience of the panel and its systematic approach to its task. Panel members should include representatives of all clinical personnel whose care is being evaluated. Both academic experts and respected practitioners are needed. It is important that they have experience in evaluating quality of care. If their product will be used nationwide, panel members should be drawn from many geographic areas and from both large and small practice settings.

The panel members provide the clinical expertise for completing the steps in Table 4.3. They are supported by panel staff who provide technical assistance. The panel staff should include expertise in quality measurement, health information management (formerly known as medical records management), nursing, and, if appropriate, computer programming. The staff analyze panel decisions for consistency and completeness and translate decisions into written formats.

An overview of suggested work procedures for the panel is given here and explained in detail below. The expert panels commonly do their work in a series of three 2-day meetings. These meetings are interspersed with mailings of materials the panel staff assemble to document work completed at the prior meeting and to prepare panel members for the tasks at the upcoming meeting. Before the first meeting, panel staff circulate to panel members materials about the guideline and about the panels tasks.

At its first meeting, the panel typically addresses the tasks of the planning phase (steps 1–4) and steps 5–9 of the development phase. Panel staff then prepare and circulate materials that document the choices made and add further details regarding the review criteria, the data sample, the data items, and the analysis procedures that comprise the performance measure.

At the second meeting, the panel reviews and finalizes the content of the proposed performance measure. Only review criteria and the criteria algorithms accepted by all panel members are included. Panel staff develop the actual instrument for performance measurement—including

abstractor manuals and training materials, and procedures for data analysis—and then conduct a pilot test that includes checks of validity and inter-rater reliability (steps 10–12).

At its third meeting, the panel reviews the detailed results of the pilot test and validity review and finalizes the instrument design. Panel staff then develop a computer data system to conduct the review according to the design.

Planning Phase

Step 1. Clarify the purpose of the performance measurement. There are many potential users of performance measures; each user must clarify the purpose of his or her review and how a performance measure will serve this purpose. Any of several events may suggest the need for the review: a regulatory requirement, patient complaints, a commitment to quality improvement, or the organization's desire to determine whether its members are following the current recommendations of their professional group. The QA/QI committee in a single organization generally seeks to measure performance initially, then takes action to improve it, and finally reassess performance to determine whether improvement has resulted.

Step 2. Identify a relevant clinical practice guideline. A clinical practice guideline is selected as the basis for a performance review, according to the

Cataract example: Step 2. Identify a relevant clinical practice guideline

For the examples used throughout this chapter, imagine that the quality assurance committee of a managed-care organization, hospital, or group practice that has many practitioners and is located in a large, populous State has chosen to study the management of functional impairment due to cataract for a quality improvement project. This topic was chosen because cataract removal is the most common surgical operation in the elderly and, if performed in the absence of appropriate indications, is the cause of unnecessary surgical risk and inconvenience as well as expense. On the other hand, if the procedure is not performed when indicated, persons who could benefit may experience decreased quality of life and loss of independence.

The example quality assurance committee identifies the AHCPR-supported guideline on cataracts in adults (O'Day, Adams, Cassem et al., 1993) as relevant to its needs. Information derived from this performance review will help clinicians and managers in the organization determine whether the care received by their patients conforms to recommended practices. If the care does not conform, the results will guide their efforts to change the delivery of care so that it is consistent with the guideline recommendations.

purpose of the review. Several guidelines may be relevant, at least in part. The committee obtains these guidelines and evaluates whether they are based on scientific evidence and were developed through a rigorous systematic process.

Step 3. Identify populations covered by the guideline. Clinical practice guidelines are written for broad patient populations. For example, the AHCPR-supported guideline *Acute Pain Management: Operative or Medical Procedures and Trauma* (Carr, Jacox, Chapman et al., 1992) addresses hospitalized patients except those with chronic pain. The guideline distinguishes specific subgroups of patients, such as children, victims of injury, and the elderly. It also addresses the care of typical surgical patients. In planning a performance measure, it is important to know which populations are covered by a specific guideline in order to select the population to be reviewed.

Cataract example: Step 3. Identify populations covered by the guideline

The committee examines the clinical practice guideline to determine which populations are addressed in the recommendations. The guideline applies to all patients with unilateral or bilateral cataracts, excluding those who are legally blind or have serious noncataract eye disease involving both eyes. The population consists of two subgroups, each of which requires somewhat different review criteria. The first subgroup consists of patients who have one or more cataracts, and not more than one eye that is legally blind or has serious noncataract disease. The second subgroup consists of patients with one or more cataracts who also have posterior capsular opacification. Since patients in the second subgroup cannot be identified at the initial workup for cataract, recommendations for their care are the same as for the first group until the additional condition is identified. At this time, these patients should have additional examinations. The committee decides to review care given to patients in the first subgroup.

Step 4. Identify guideline recommendations and draft the medical review criteria. A clinical practice guideline can usually be divided into several major areas, which may reflect the timing, sites, or techniques for sequences of care. Examples of these broad areas are initial evaluation of a patient's symptoms; testing to confirm a diagnosis; treatment; and followup. Other sections may concern outcomes and administrative requirements. The committee first identifies the major sections of the clinical practice guideline that are relevant to its purpose.

AHCPR-supported guidelines are issued in four versions: a clinical practice guideline, a quick reference guide, a patient's guide, and a guideline report. Of the four versions, the quick reference guide may be most suited to helping the committee identify the major sections for which it will measure performance.

Cataract example: Step 4. Identify guideline recommendations . . .

The guideline for cataract management can be divided into seven major sections:

1. Initial evaluation.
2. Counseling and decision regarding surgery.
3. Nonsurgical management.
4. Preoperative management.
5. Operative procedures and indications for hospitalization.
6. Postoperative care.
7. Rehabilitation.

The committee then chooses the specific clinical decisions and actions recommended in the guideline that are relevant to its purpose. Since reviewing all the guideline recommendations is seldom financially feasible, the items with the greatest impact on patient health or greatest relevance to obtaining value for money are usually selected for review. Impact is considered great when an issue affects a few patients severely or affects many patients. For example, failure to ask all patients about drug allergies may have serious results for the few who will experience anaphylactic shock if given drugs to which they are allergic; on the other hand, failure to check antibiotic sensitivity for organisms creating urinary tract infections seldom causes a life-threatening situation but may prolong the illness and worsen the discomfort of large numbers of patients.

Recommendations for suitable use of dangerous or expensive technologies are important to include. Recommendations concerning inexpensive and noninvasive tests and treatments are important because unneeded tests and treatments expose patients to unnecessary risks while consuming resources, whereas failure to perform needed tests and treatments may prolong illness.

If the purpose of review relates to the performance of only one type of clinician, guideline recommendations for activities performed exclusively by other types of clinicians should not be included. For example, a review of anesthesiologists' conformance to the AHCPR-supported guideline on acute pain management (Carr, Jacox, Chapman et al., 1992) might not include behavioral techniques for pain control if nurses are the ones who educate patients about such techniques.

A lack of data to determine guideline conformance may also eliminate certain sites of care from the review process. For example, ambulatory records may be more likely to contain the information needed to evaluate guideline

conformance in the management of urinary incontinence. However, the committee may decide to include even those care processes known to be poorly documented by many clinicians. Feedback regarding the range of performance rates across the group of clinicians being studied can then be used to demonstrate that better documentation is both needed and achievable.

A clinical practice guideline recommendation generally takes the form of advice such as, "Use this drug for patients with this clinical condition." A medical review criterion states the action that indicates conformance to the guideline—for example, "This drug was used for patients with the specific clinical condition." For each guideline recommendation to be reviewed, the committee drafts a medical review criterion plus any related acceptable alternatives and exclusions as described in the guideline.

Cataract example: Step 4. . . . and draft the medical review criteria

From the major areas of the cataract guideline, the committee decides that the initial evaluation is the most relevant area for its first attempt at guideline-related performance measurement. Note that initial evaluation is only the first of seven sections of the guideline listed in the preceding shaded box. Initial evaluation is the phase of patient care that can be performed by the largest number of clinicians and that determines the need for surgery. After carefully considering the purpose of its reviews, the committee lists the following draft criteria and the acceptable alternatives that are suggested by the guideline:

1. A medical history is taken, including patient report of level of disability.
2. A social history is taken, including patient report of preferences regarding surgery.
3. A general physical exam and testing are done.
4. A complete ophthalmological exam is done.
5. A Snellen visual acuity test is done.
6. No other vision tests are done. (Acceptable alternative: A glare test may be done if the patient's vision is only slightly impaired but the patient complains of glare.)

When multiorganizational performance measurements include clinicians with different backgrounds or different practice settings, some guideline pathways may be selected for criterion development that would not be relevant to a single organizational setting. For example, a hospital quality assurance committee may omit from its review criteria guideline recommendations that have already been formally adopted and institutionalized as clinical policies. However, in a nationwide study it is advisable to

include all the guideline recommendations that relate to the purpose of review, since some hospitals may not yet have adopted all such practices.

Development Phase

Step 5. Identify clinicians and sites of care. The organization is already committed to reviewing the care that is given in its own sites. Determining who is included in the review narrows choices for the patient sample and the final form of the review criteria. For example, in assessing compliance with the AHCPR-supported acute pain management guideline, a hospital committee may limit its review to patients having surgery in orthopedic surgical units, thus excluding other surgical units. An HMO committee reviewing adherence to the American Academy of Pediatrics guideline for childhood immunizations may exclude a center located in a county that gives free immunizations in the public schools.

Cataract example: Step 5. Identify clinicians and sites of care

Initial evaluation of patients with cataracts is conducted in ambulatory settings, such as clinicians' offices and clinics. Physician and nonphysician primary care practitioners may do part of the initial evaluation and then decide whether to refer patients to an ophthalmologist, who conducts all or only part of the evaluation. The committee decides to review the performance of primary care clinicians, ophthalmologists, and their support staffs. The clinicians include family physicians, general practitioners, internists, physicians' assistants, nurse practitioners, osteopathic physicians, ophthalmologists, and optometrists.

In large-scale reviews covering many geographic areas and many types of organizations, identifying clinicians and sites of care for review requires careful thought. For example, the AHCPR-supported guideline Urinary Incontinence in Adults *(Diokno, McCormick, Colling et al., 1992) concerns care delivered in ambulatory settings, hospitals, and nursing homes. These settings differ so much in terms of their clinician and patient populations and in availability of data and data formats that performance measurement must be adapted specifically to each type of site.*

In choosing to measure conformance with a particular guideline recommendation, the panel may have automatically determined the site of care in which review will occur. For instance, if the type of surgery performed for urinary incontinence is the item of interest, and this surgery is usually performed in a hospital, then review will focus on the hospital setting. However, if bladder training is the item of interest, conformance to the guideline could be assessed in clinicians' offices, in nursing homes, or in hospitals. If the panel wishes to compare conformance with the bladder training recommendation at all three types of sites, the panel

(guided by technical staff and consultants) must plan in detail how to abstract data from the three different types of data sources in such a way that a fair comparison can be made.

Step 6. Define case sample and case sampling period. The terms most frequently used in the process of defining the case samples and the case sampling period are defined here:

- **Case sampling period.** The time period during which a case is considered eligible for inclusion in the denominator of a performance measure.
- **Denominator event/state, index event/state.** The event or health state that defines a patient's eligibility for inclusion in the denominator group for a performance measurement.
- **Denominator for a performance measure.** The sample of cases that will be observed to determine conformance to medical review criteria.
- **Exclusion.** Characteristics or conditions that make cases ineligible for review by a specific performance measure or by a specific criterion within a performance measure.
- **Index event/state.** See **Denominator event/state**.
- **Numerator for a performance measure.** An event specified in a medical review criterion as evidence of guideline conformance.
- **Performance rate.** A measurement produced by using a performance measure, providing a quantitative evaluation of events related to patient care. A performance rate results when the numerator for a performance measure is divided by the denominator for that measure.
- **Sample.** The subset of a population or the group of cases to whom a performance measure will be applied in order to assess rates of conformance to a clinical practice guideline.
- **Time window.** The time period following an index event during which a case is "observed" for evidence that the care did or did not conform to medical review criteria. In other words, the interval in which it must be determined whether or not a numerator event took place.

To ensure the validity of the performance measure, the right event or state to define the denominator of a performance measure must be chosen. Furthermore, the medical review criteria that define the numerator of the performance measure must be applied to the denominator group. For example, conformance to a guideline for breast cancer *screening* is measured in a general population of women, excluding those already known to have breast cancer, but conformance to a guideline for breast cancer *management* is measured in women already known to have breast cancer.

To define the case sample, the committee specifies in detail the characteristics by which cases will be selected for inclusion in the denominator. Selection may be based on an event. For instance, a hospital committee reviewing conformance to the AHCPR-supported guideline for acute pain management (Carr, Jacox, Chapman et al., 1992) decides in step 5 (Table 4.3) to focus on the orthopedic surgery department; in step 6, therefore, the committee selects orthopedic surgery patients, and an orthopedic operation as the index event.

Case selection may also be based on patient state. For instance, in the preceding example, the committee restricts its review by excluding patients who are substance abusers, since the guideline on acute pain does not apply to such patients. Further, the committee might restrict the patient sample to those having surgery during or after 1992, the first full year after release of the AHCPR guideline. In this way, the committee measures conformance to the guideline only after orthopedic surgeons could have known what the guideline recommended.

As a second example, an HMO committee reviewing conformance to the American Academy of Pediatrics guideline for childhood immunizations selects children aged 2 and older for the case sample, because the guideline recommends that the initial immunization series should be completed by age 2. Thus, being a child aged 2 or older is the index state. The committee also decides to include only the infants who have remained continuously enrolled since birth in order to test the plan's performance for the infants for whom it has been solely responsible. (This decision excludes infants who failed to receive immunizations while their parents lacked health insurance or who belonged to other plans. Studying this group tests a different aspect of performance, namely, whether the plan succeeds in catching up with the immunization schedule for infants who are enrolled in the plan after they are born.) Furthermore, the committee limits the case sampling period for infants to those born during the period 2–4 years before the study. In this way, the committee restricts its focus to the care given to infants during the previous 2 years. Such thoughtful specification of inclusion and exclusion rules ensures that the case sample used for the denominator of the performance rate is uniformly relevant to the purpose of measurement.

Although a practice guideline is written broadly to include many manifestations of a condition, measuring conformance to a guideline must be narrowly focused in order to reduce error and thus enhance the validity of the measurement. Patient subgroups for whom the guideline makes different recommendations must be identified as separate samples, and separate criteria sets must be developed for each such subgroup. For example, the broad group of patients with urinary incontinence is split into subgroups of patients: patients with incontinence due to central nervous system disorders, patients with urinary tract infection, males with prostatic hyperplasia, females with stress incontinence, and females with urge incontinence. The committee may have to exclude from the sample patients with comorbidities that indicate a different course of management; for example, management of urinary incontinence is different for patients with diabetes, and treatment of anemia requires a different approach in patients who are pregnant versus those who are not. In addition, current

medications may dictate exclusions from a patient sample; for example, guidelines differ for insulin-dependent diabetic patients versus non-insulin-dependent diabetic patients. Prior history is also a consideration: management of benign prostatic hyperplasia is different for patients who have had prostate surgery; treatment recommendations are different for patients with different numbers and types of vessels affected by coronary artery disease. When there are inclusions or exclusions based on diagnosis or procedure, careful selection of standardized diagnosis codes (e.g., ICD-9-CM) and procedure codes (e.g., CPT4) helps to define the sample precisely.

Because guideline recommendations change over time, specifying the case sampling period is important. For instance, a committee may reasonably expect less conformance from clinicians in regard to care they gave before the guideline was published. To measure guideline conformance, therefore, the committee may set the case sampling period to include patients who became eligible for care after the guideline became available.

Cataract example: Step 6. Define case sample and case sampling period

The committee decides to limit the review to patients aged 65 and older. Patients with concomitant severe eye disease are excluded because their treatment differs from that of most cataract patients. Also excluded are institutionalized patients and patients who cannot communicate. The case sampling period is set as 1 year before initiation of the review.

Thus, the patient sample is defined as all males and females older than 65 who were seen in an ambulatory setting within 1 year before the date of review, who complain of vision problems, and who are given a diagnosis of cataract, with the exclusions named above. The committee decides on the following two-stage sampling method:

1. Find the records of all patients over 65 with cataract that are coded for a visit within the past year and not coded for a specified list of concomitant serious eye diseases, of which the most common are glaucoma and diabetic retinopathy.

2. Examine the records for the first mention of a cataract. If the first diagnosis is earlier than 1 year ago, eliminate the case from the sample. Also exclude patients who have already had bilateral cataract surgery or other causes for exclusion not detected in the first stage of sampling by diagnosis codes.

Regional differences in patient populations are a concern for panels planning multiorganizational measurements. If the panel elects to review care in a variety of geographic locations, provision is made for differences in the prevalence of guideline-related and comorbid diseases among the patient populations in those areas. Because the prevalence of this disease in the local population affects the probability that a patient with positive diagnostic

test for the disease truly has the disease, a clinician may justifiably modify a practice guideline to adjust for the probability that a patient in the local population truly has the guideline-related disease. For each patient characteristic for which substantial regional differences exist and for which guideline recommendations might reasonably be modified, the criteria must stratify the review, using "branching logic"—that is, citing two pathways, one for patients with the characteristic and the other for patients without it. Alternatively, allowance for patient differences is made in applying comparative standards to performance rates (see Appendix C).

In a multiorganizational study, data collection may spread over months or years. The panel considers carefully the logistics of data collection in regard to setting the case sampling period. When measuring conformance to a newly issued clinical practice guideline, the findings could be misleading if, for example, the case sampling periods for East Coast patients and West Coast patients were in two different years. Similarly, because of seasonal variations in illnesses afflicting patients and in staffing patterns in sites of care, rates based on a 12-month case sampling period are not comparable with rates based on an 18-month case sampling period.

When the patient sample is drawn from an administrative data base, it is important that all patients be continuously eligible for inclusion in the data base during the entire case sampling period and time window. For example, if the Medicare National Claims History File is being used, the patients sampled for a performance measurement must be 65 years old or older at the beginning of the time window. For review of managed care organizations, patients should be continuously enrolled members of the health care plan during the time window.

To keep performance measurement within reasonable bounds of time and cost, the panel must design samples that include relatively small numbers of patients who are representative of very large and heterogeneous populations. Sound methods for sampling are important in achieving good comparability among the patients whose care is being used to measure guideline conformance. Methods for selecting appropriate random samples and stratified random samples are described in Appendix C.

Step 7. Identify data source. The terms defined below are those most commonly used in discussions of data types:

- **Outcome data.** Data describing a patient's health status.
- **Process data.** What is done to, for, or by patients as part of the delivery of care, such as the performance of a test or procedure.
- **Structural data.** Information about organizational facilities, equipment, policies, and procedures; for example, a hospital policy for patient-controlled analgesia.

Evidence for conformance to clinical practice guidelines may be contained in several data sources. Traditionally, the process and outcome data used for performance measurement come from patient records maintained by providers of health care such as hospitals, HMOs, physicians' offices, home care agencies, and nursing homes. An organizational committee may choose additional sources of data for specific items, such as records kept by the pharmacy, laboratory, and radiology facilities. Administrative data such as registration and claims files can provide patient demographic information and documentation of reimbursable services. For instance, conformance to a recommendation in the urinary incontinence guideline about frequency of changing pads for incontinent patients can be measured by counting the number of incontinence pads charged to a patient. Less formal data sources include log books that note the referrals made, appointments scheduled, and appointments not kept.

In addition, the committee can survey patients to elicit reports of satisfaction with and facts about the services they receive. Patient questionnaires are particularly useful when the committee seeks information known only to the patient, such as whether the patient understood all relevant treatment options. Less commonly, the committee may collect data for a particular performance measurement by surveying clinicians, by incorporating new data collection into patient care routines, or by directly observing patient care. For instance, conformance to guideline recommendations concerning patient counseling can be assessed by observing counseling sessions.

Some review criteria require structural data about health care facilities, such as the existence of certain policies and the capacity to perform certain functions. The committee may find this information in policy documents and through surveys of administrators and clinicians. An example of such a criterion, derived from the AHCPR-supported guideline on acute pain management (Carr, Jacox, Chapman et al., 1992), is that "the hospital should have a formal mechanism to evaluate pain management." Hereafter, unless otherwise specified, the term *data source* as used in this document applies to patient records, since patient records are a common source of data for assessing practice guideline conformance.

The review committee also defines the time window, the time interval necessary to complete the actions specified by the guideline recommendations and, therefore, the time interval to be searched in the data source. For example, in the guideline on acute pain management, many recommendations pertain to specific intervals within the postoperative period: the time window for applying the related review criteria must cover the same time intervals. The data source is searched for the entire time window in order to determine whether care conformed to the practice guideline.

The time window is tied to the date of the index event. It may be set as a point in time (a guideline recommendation should be followed at a given time at or after the index event); a timeframe (a guideline recommendation should be followed within a given period before or after the index event); or an

episode of care (a guideline recommendation should be followed during an episode of hospitalization, illness, or outpatient treatment). An episode of care has appeal as a natural unit of clinical care, but it is imprecise because of the difficulty in defining its beginning and end. In addition, the length of episodes can vary substantially for different patients. Here are two examples of time windows for specific medical review criteria: "An influenza vaccination should be given to all eligible patients at any clinic visit between September and February." "A repeat hematocrit should be determined to confirm anemia within 3 weeks of an initial low hematocrit."

Cataract example: Step 7. Identify data source

Patient records from the clinicians and sites of care specified in step 5 are chosen as the data source for documenting criteria conformance. Patient records are an appropriate data source specifically for the part of the guideline identified by the committee as being concerned with initial evaluation. However, if the committee wished to evaluate conformance for the part of the guideline identified as being concerned about counseling and decision regarding surgery, the physician's perceptions, even if written in the patient record, are not sufficient. To determine whether patients understood any counseling and felt involved in the decision regarding surgery, it is necessary to ask patients themselves through questionnaires or interviews.

The panel developing multiorganizational performance measures concerns itself with the comparability of patient care data from the various settings where guideline conformance will be reviewed. Organization of patient records and conventions for documenting patient care vary from region to region and from clinician to clinician. Data collection must be tailored to local idiosyncrasies in such a way as to ensure that data items are as equivalent as possible. For example, if followup care after surgery is reviewed with ambulatory care records as the data source, some records may include hospital discharge summaries while others do not. The panel decides whether data items in discharge summaries are essential for performance measurement; if so, any discharge summary that is missing from an ambulatory record must be separately obtained. Failure to identify such discrepancies in the data at different care sites and providers leads to unfair comparisons of guideline conformance.

Step 8. Write medical review criteria, specifying acceptable alternatives and time window. The technical terms used in this section are defined below:

- **Acceptable alternative.** A common and legitimate reason for not conforming to practice guideline recommendations; for example, the clinician recommended a treatment according to the guideline, but the patient refused. Acceptable alternatives are specified explicitly when writing review criteria, whether the alternatives have been stated

explicitly in the practice guideline or merely implied. "Acceptable alternative" is also the name of the status assigned to a criterion during a review if documentation is found for a defined acceptable alternative to the criterion.

- **Algorithm.** A rule of procedure, or set of instructions, containing conditional logic for solving a problem or accomplishing a task. Guideline algorithms relate to recommendations for patient care (Gottlieb, Margolis, and Schoenbaum, 1990; Hadorn, McCormick, and Diokno, 1992; Margolis, 1983). Criteria algorithms concern rules for evaluating criteria conformance. Algorithms may be expressed in words only or in diagrammatic form (Margolis, 1992).

- **Criteria set.** A series of criterion statements linked together because they all apply to the same patient sample.

- **Criterion status.** The category to which a case is assigned by application of a criterion. For example, the case that meets a criterion is assigned the status "met"; a case that meets an *acceptable alternative* to a criterion (see definition above) is assigned the status "acceptable alternative." If a criteria set incorporates branching logic (i.e., a certain criterion applies only to a defined subgroup of cases), cases not within that subgroup are assigned the status "not applicable" for that criterion. If data needed to determine whether a case met a criterion are not found in the selected data source, the status "not reviewable" is assigned. A case that does not fit any of the above categories is assigned by default to the status "not met" (i.e., the care given to the case did not conform to the practice guideline).

- **Medical review criteria.** Systematically developed statements that can be used to assess specific health care decisions, services, and outcomes. Each criterion derived from a guideline recommendation is used to determine whether the case being reviewed conforms to a particular recommendation in the guideline. A status is assigned to each criterion to reflect the care given.

In drafting criteria (step 4), the committee incorporates the exclusions from the patient sample and the acceptable alternatives to review criteria that are identifiable in the guideline. In step 8, as the criteria are specified in detail, the committee adds exclusions and acceptable alternatives to refine the validity of the performance measure for the committee's purpose. The main concern is to reduce the number of false negative identifications of nonconformance to the guideline that occur when the guideline cannot or should not be applied to a case. The committee should be careful, however, not to exclude cases or allow acceptable alternatives that undermine the intent of the guideline. The following examples illustrate reasonable exclusions and acceptable alternatives.

Cataract example: Step 8. Write review criteria, specifying acceptable alternatives and time window

The committee discusses each item to be abstracted for evaluation of criteria compliance and completes the worksheet. If the committee does not include an ophthalmologist and a health information management professional, these individuals are consulted regarding the specification of the data required for evaluating compliance with the criteria and in the decision rules for obtaining the relevant data. The AHCPR-supported cataract guideline (O'Day, Adams, Cassem et al., 1993) does not specify a time period within which the initial evaluation ought to be performed; however, the committee, in consultation with the ophthalmologist, decides that it will evaluate whether the first diagnosis of cataract was investigated within 6 months. They also explore the justifiable reasons for failure to investigate this finding. In the fourth column, the committee defines the specific items of "medical history," "social history," "ophthalmological examination" that the abstractor may accept as evidence of an "appropriate" history or examination. These reasons, the acceptable alternatives to criteria conformance, and additional definitions of data items are shown on the completed worksheet in Table 5.2.

After the criteria development worksheet has been completed by the committee, the criteria may be written in narrative form for dissemination to clinicians, reviewers, and other interested parties.

Narrative form of criteria for initial evaluation and testing

The example of criteria in narrative form shown below matches the criteria shown on the last four rows of Table 5.2. The test criteria shown in narrative form here appear in the last three rows of Table 5.2 and in the algorithm form in Figure 5.1.

Ophthalmological examination criterion:

A complete ophthalmological exam should be done by an ophthalmologist within 6 months of the diagnosis of cataract. The exam should document positive or negative findings of the following:

- Appearance of the lens.
- Appearance of the cornea.
- Appearance of the retina and/or macula.
- Intraocular pressure.

Failure to meet the criterion is justified if the patient refuses the exam.

Test criteria:

1. A Snellen test of visual acuity should be done within 6 months of the diagnosis of cataract. Documentation of best corrected vision

(continued from p. 46)

should be present for each eye individually and both eyes together.

2. A glare test should not be done unless the patient complains of glare and has a visual acuity of 20/40 or better as measured by the Snellen test.

3. No other vision tests should be done for the initial preoperative evaluation of cataract. These tests include, but are not limited to, the following: contrast sensitivity, potential vision, specular photographic microscopy, formal visual fields, fluorescein angiography, external photography, corneal pachymetry, B-scan ultrasonography, electrophysiological tests.

Some exclusions remove patients altogether from the sample for review. For instance, it is reasonable for patients with urinary incontinence who are also terminally ill not to be evaluated and treated in the same way as patients who are not terminally ill. "Terminal illness" may be a reason for exclusion from the patient sample for review (i.e., from application of all criteria in the set for review of management of urinary incontinence). Exclusions may also apply selectively to particular criteria that are contingent on a given patient state. For example, for patients with symptoms of benign prostatic hyperplasia who also have hematuria, an intravenous urogram may be indicated, although it is not recommended for patients without hematuria. The committee writes a contingent criterion that excludes patients without hematuria and applies only to patients with hematuria.

An acceptable alternative to a review criterion is a way of allowing for different options that may apply to the patients eligible for a given criterion. For instance, the AHCPR-supported acute pain management guideline (Carr, Jacox, Chapman et al., 1992) recommends postoperative use of nonsteroidal anti-inflammatory drugs (NSAIDs) for all general surgical patients. The guideline does not mention allergy or adverse reactions to NSAIDs, but since some patients are indeed allergic to or have had adverse reactions to NSAIDs, it is obviously good care not to give such patients NSAIDs. In such cases the criterion will not be met, but there is an acceptable reason. "Allergy to NSAIDs" is an "acceptable alternative" to the review criterion testing for NSAIDs use.

Note that timeliness of performance of an action recommended by a guideline, although not mentioned in the guideline, may be implied. If the guideline recommends a test, with action to follow upon the result of the test, it implies that the test and resulting action must occur within a time frame that permits the action to improve patient outcome.

The committee does its work using a three- or four-column format, shown in Table 5.1. (Use of a three-column format is first described in Jacobs, Christoffel, and Dixon, 1976.) Table 5.2 is drawn from the AHCPR-supported

Table 5.1. Sample of a partially complete criteria development worksheet (cataract example)

Data elements	Acceptable alternatives	Data sources	Abstracting instructions/explanatory notes
Medical history	History recorded by other physician is resummarized in record	Patient record	
Social history	History recorded by other physician is resummarized in record	Patient record	
General physical exam and testing	Examination recorded by other physician is resummarized in record	Patient record	
Complete ophthalmologic exam	Patient refusal	Patient record	
Snellen visual acuity test	None	Patient record	
No glare test	Complaint of glare *and* Snellen 20/40 or better	Patient record	
No other vision tests	None	Patient record	

Table 5.2. Sample of a complete criteria development worksheet with data items specified (cataract example)

Data elements	Acceptable alternatives	Data sources	Abstracting instructions/explanatory notes
Medical history	History recorded by other physician is resummarized in record	Patient record: visit note, preop note	≤ 6 months after diagnosis; presence/absence of diabetes, cardiovascular disease, respiratory disease, renal disease, major neurological disease
Social history	History recorded by other physician is resummarized in record	Patient record: visit note, preop note	≤ 6 months after diagnosis; document functional status: independence, ability to perform activities of daily living (ADL), driving, reading, television, work; mention of living arrangements or assistance available after surgery; psychosocial or economic problems
General physical exam and testing	Examination recorded by other physician is resummarized in record	Patient record: visit note, preop note	≤ 6 months after diagnosis; review of systems, must include mention of heart, lungs, renal, and neuromuscular; complete blood count
Complete ophthalmologic exam	Patient refusal	Patient record: visit note, ophthalmologist consult note, preop note	≤ 6 months after diagnosis; must be done by ophthalmologist; must include appearance of lens, cornea, and retina/macula, and intraocular pressure
Snellen visual acuity test	None	Patient record: visit note, ophthalmologist consult note, preop note	≤ 6 months after diagnosis; must document best corrected vision: each eye separately and both together
No glare test	Complaint of glare *and* Snellen 20/40 or better	Patient record: visit note, ophthalmologist consult note, preop note	
No other vision tests	None	Patient record: visit note, ophthalmologist consult note, pre-op note	Includes contrast sensitivity, potential vision, specular photographic microscopy, formal visual fields, fluorescein angiography, external photography, corneal pachymetry, B-scan ultrasonography, electrophysiological tests

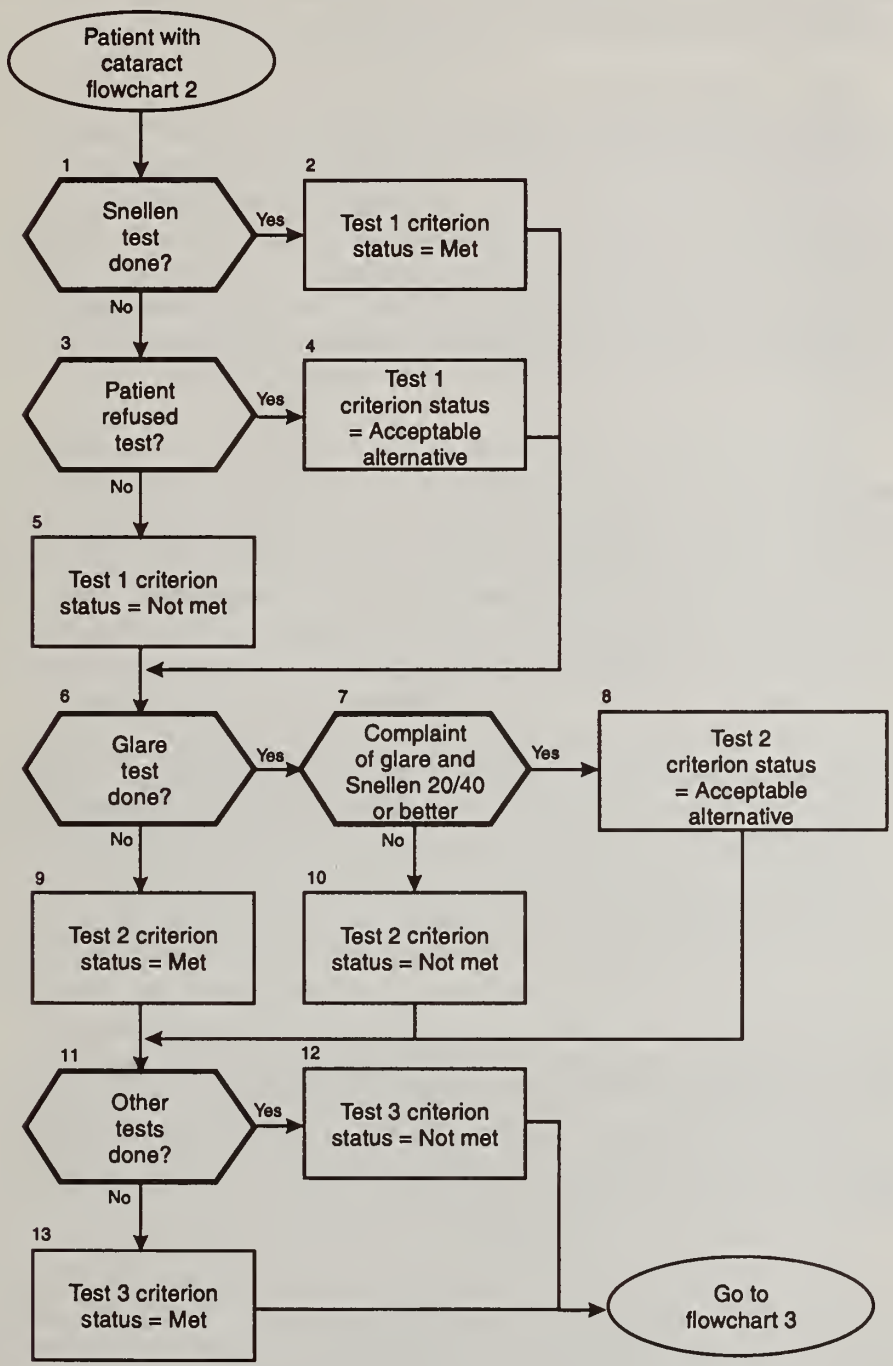
NOTE: The items shown here as an example were recommended as usual practice by a single practicing ophthalmologist.

guideline on cataracts (O'Day, Adams, Cassem et al., 1993). As used here, each row relates to a criterion statement. The first column is used to define an element of patient care that provides evidence of conformance to a criterion. In the second column the acceptable alternatives are listed; the third column specifies the location where the data item should be sought; the fourth column provides explanations of the criterion and data items and preliminary abstracting instructions. The set of criteria that applies to a defined patient sample is recorded in order down the rows of the sheet.

Step 9. Specify data items and data rules. This stage of development of a performance measure includes detailed definitions of the needed data items, design of a data collection form, and the writing of instructions or a manual of procedures for abstractors. Data items are needed to delineate the patient/event sample and any exclusions from the sample (to identify the patients/events counted in the denominator). Other data items are needed to identify the patients/events that conform to the guideline (those patients/events to be counted in the numerator). For many criteria, the data needed are quite evident: "Repeat hematocrit is done within 3 weeks" requires finding any hematocrit and the date it was done. For other criteria, the data required must be specified further. For example, a criterion drawn from the AHCPR-supported urinary incontinence guideline (Diokno, McCormick, Colling et al., 1992) states "treatment was appropriate to the urinary incontinence diagnosis." All possible urinary incontinence diagnoses (urge, stress, mixed) and the treatments appropriate for each (surgery type, specific medications, behavioral interventions) must be identified. The data items that document these elements of care are abstracted from the patient record so that the criterion may be applied to determine whether treatment was appropriate.

The committee uses the fourth column on the worksheet to define for each criterion the data items that are accepted as evidence that the particular criterion is met or that an acceptable alternative exists. This column contains definitions of terms used in the criterion statement. For example, in the AHCPR-supported acute pain management guideline (Carr, Jacox, Chapman et al., 1992), certain actions are recommended if the patient has a history of adverse effects from pain medication: the committee defines what evidence in the patient record justifies a conclusion that the patient has had adverse effects. The committee may also define what sections in the data source must be searched before concluding that a relevant data item does or does not exist. For instance, to determine whether "respiratory depression" exists, the committee may require an abstractor to search for such a diagnosis in the physician notes and also to search the nursing notes for relevant comments together with a record of an actual respiratory rate that meets the committee's definition for respiratory depression. For some data items, the committee may decide that certain sections of the data source should not be used. For example, to establish a diagnosis of osteoarthritis in order to monitor the treatment given, the committee instructs abstractors to search the physician notes for mention of this diagnosis, but not to accept mention of asymptomatic osteoarthritis discovered as a minor or incidental finding on an X-ray report. Most individuals

Figure 5.1. Flowchart version of the algorithm for assessing conformance to review criteria (cataract example, section on vision tests done on initial evaluation)



NOTE: Flowcharts 2 and 3 not shown.

acquire minor degrees of osteoarthritis as they age. Asymptomatic, minor disease does not justify the expensive and potentially risky tests and treatments that are indicated for more serious disease; therefore, the asymptomatic patients, ascertained incidentally in X-ray findings, are excluded from measurement of guideline conformance for osteoarthritis.

The committee also adds further details in columns three and four of the worksheet. In the fourth column, the committee defines the time window for each criterion, i.e., for what period of patient care must the data source be searched before concluding that a relevant data item does not exist? In the fourth column, also, the committee may add further instructions for the data sources to be searched for evidence of conformance to each criterion (see Table 5.2). In this example, by excluding all vision tests other than visual acuity and glare tests, the committee ensures any patient receiving such other tests receives individual case-based review.

For small, informal studies of guideline conformance, reviewers can be trained to use the worksheet directly to guide review. When the review is conducted, the reviewer works down the rows of the criterion worksheet to assign a criterion status to each case included in the patient sample that makes up the denominator of the performance measure. Consider the example of a criterion requiring that a mammogram be done. If the mammogram was done, the criterion status is "met." If the reviewer finds evidence of a defined acceptable alternative, the criterion status is "acceptable alternative," e.g., the patient refused the mammogram or is suffering from a comorbid terminal illness that makes mammogram screening irrelevant, or receives gynecological care from another clinician who is not included in this performance review. If the patient is not eligible for the criterion, the criterion status is "not applicable." For instance, if the patient has already had bilateral mastectomies, the mammogram criterion is not applicable. If the committee defines the evidence that must be obtained in order to assign criterion status, and that evidence is not found in the patient record, the criterion status is "not reviewable," e.g., the mammogram care should occur in the prior year but the patient was first seen only 3 months earlier. If the committee defines the absence of documentation of a recommended action as evidence that it did not occur, and no documentation of the recommended action is found, the criterion status is "not met," e.g., although the patient was under care throughout the prior year, there is no evidence of a mammogram having been done.

It is tempting to collect more data than are needed to assess conformance to the criteria. Once data sources have been accessed, some committee members may argue for collecting data to answer additional questions of interest to them. For example, how many of the urinary incontinence patients have a college education? How many patients receiving surgery for benign prostatic hyperplasia are covered by insurance? Unless resources for review are abundant, the committee resists including additional questions because the cost of abstracting data is directly related to the number of data elements abstracted, recorded, and analyzed. Committee staff advise the committee to adopt a frugal

approach to data selection. Data are generally included in a performance measure only if they have one of the following functions:

- *Help assess the criterion.* For example, to assess whether an abnormal hematocrit was rechecked within 21 days, data to be abstracted are a record of the test and of its date.

- *Help follow decision rules for evaluating criteria compliance.* For example, if there are different criteria for men and women, it is necessary to know patient gender.

- *Define exclusions or acceptable alternatives.* For example, in reviewing care given to postoperative pain patients, it is necessary to exclude patients from the numerator who are drug dependent and to have a data item that notes such drug dependency, because different criteria apply to such patients.

- *Identify gaps in knowledge pertaining to the clinical practice guideline.* Other data that may not be part of a guideline may be desired to obtain knowledge of current practice. For instance, the committee may decide to collect data on a new drug that came into use after the guideline was written, including how it is used and its effects. To control costs, such additional items are not collected if the question has been addressed in other studies.

Step 10. Draft data collection forms and procedures. Careful design of data collection forms and procedures is critical for ensuring both the reliability and validity of a performance measure. With attention to detail, the committee, guided by its technical support staff, develops instruments that ensure consistent and relevant data collection. In order to protect rights of individuals to privacy, the data collection forms use codes instead of names to identify patients and clinicians.

The data abstraction form is formatted so that it promotes accuracy in filling in the blanks, limits the likelihood of missing data items, and makes it possible later to collate the data by hand or enter it into a computer file quickly and accurately. In general, the questions on the form follow the order in which the information appears in the patient record or other source of data. If data are to be found in more than one place in a patient record or in more than one document, the questions are grouped according to the source, or there are different forms for each source. For example, when abstracting from the ambulatory record, the first section of the form collects information from all visit notes, and subsequent sections address laboratory tests, radiology reports, and consult notes, in that order. Finally, the last section or a separate form is used for data obtained from the patient's hospital record or discharge summary.

Abstraction forms may be accompanied by written instructions or decision rules for abstractors. These instructions and rules are derived from the

information that was written onto the criteria development worksheets. A summary of the same instructions may also appear on the abstracting form to facilitate data abstraction.

Variation in the procedures for data collection is a source of unreliability in quality measurements. The methods and specific instructions for data collection are, therefore, especially important in multiorganizational reviews to ensure consistency of data collection among the many abstractors and performance sites. Limiting measurement error requires rigorous quality control of data at every step including abstraction, entry, and processing.

To facilitate data quality control and subsequent analysis in large projects, a computer system is built, into which data from paper data abstraction forms are entered. For very large projects, economies of scale justify the expense of building an interactive computer system for direct entry of data from paper records or electronic data sources. The computer screen in such a system prompts the abstractor to look for a data item, giving precise instructions above the item, and when this is entered, prompts for the next item. This speeds up the data collection process while reducing opportunities for errors in data transcription.

In designing a computerized data entry system, features are incorporated to minimize errors in data entry. To facilitate data entry, codes requiring few keystrokes are recommended. Numeric codes are preferred to letters because they permit rapid data entry using the numeric key pad. Whenever possible, codes should have consistent meaning from question to question, enabling the abstractor to memorize codes. The number of characters for each field is specified, the fields are aligned and right-justified, and no blank spaces are permitted (see Figure 5.2 for an example). A leading zero is inserted when a single digit number is abstracted into a two-character field. Instead of leaving blanks for missing data, the code "9" or a combination of "9"s is commonly used. Similarly, a code is devised for data not applicable or not required for the question (frequently "8"). When the responses on the form lead to branching, for example, "If no, go to question 15," the skipped questions are filled in either by the computer or the data abstractor with the code for "not applicable." Finally, when a question lists a selection of choices, but the list is not exhaustive, there is a code for "other" (see Figure 5.2).

A set of written data collection procedures, or abstractor's manual, is essential to multiorganizational performance reviews. The abstraction manual contains examples of the appearance of the data sources being used and examples of properly completed abstraction forms. These examples demonstrate the most common situations that are encountered in the patient records, such as branching or missing data. If data are being entered directly into a computer data base, in addition to a clearly written manual for the abstractor, there are "Help" screens incorporated into

Figure 5.2. Sample of abstraction form features (cataract example)

- Blank line for each character
- Align data fields

Patient study code	__ __ __ __
Date of birth	__ __/__ __/__ __
Hct done (1 = Y, 2 = N)	__
Hct result	__ __.__

- Consistent meaning for codes
- Include code for "other"
- No blanks allowed

1. Was biopsy normal? __
   - 1 - Yes
   - 2 - No
   - 8 - Not applicable
   - 9 - Not available

2. Type of referral __
   - 1 - Surgery
   - 2 - Gyn
   - 3 - Endocrine
   - 7 - Other
   - 8 - Not applicable
   - 9 - Not available

the data entry program to display the instructions on-line. If a computerized data entry system is used, the manual has illustrations showing the appearance of the computer screens.

Step 11. Device analysis procedures. During the review analysis, data for each case in the patient sample will be compared to each criterion (e.g., each row shown in Table 5.2) in order to assign as its criterion status one of the following five options: "met," "acceptable alternative," "not applicable," "not reviewable," or "not met."

If a performance measure is applied to large numbers of cases, an orderly analysis procedure is developed to promote reliability and efficiency of review. To design this analysis procedure, the panel staff analyze the relationships between the criteria statements and acceptable alternatives by drawing a criteria algorithm flowchart. This describes an efficient pathway for each case to follow through the criteria set. The criteria algorithm provides explicit step-by-step instructions for the decisions that a data abstractor in an organizational review may perform mentally. By making each decision and its place in the sequence of review explicit, the algorithm ensures uniform application of the criteria to all cases in the sample. Figure 5.1 shows the algorithm flowchart related to the criteria for vision tests shown in Table 5.2. Notice that the criteria algorithm improves upon the form of the criteria given in words alone by specifying unambiguously how each criterion status is derived for each criterion. By creating this algorithm flowchart, the panel staff may uncover any gaps and redundancies in the logic underlying the form of the criteria written in words only. The panel is informed of any such problems at its next meeting and asked to provide a solution. Appendix D describes a method for constructing algorithm flowcharts.

In building a computerized system for performance measurement, there are two options for applying the review criteria to the patient care data. These options have major implications for the time and expense of developing and operating the review system and for the qualifications of the personnel needed to develop and operate the review system. In option A, the abstractor reviews the data items selected by the panel, applies each criterion mentally, preferably following an appropriate algorithm, to define criterion status, and records or enters into the system the criterion status. The data system for this approach is very simple, but there is a risk of human error introduced when abstractors interpret and apply the criterion. In option B, the abstractor records or directly enters the data items selected by the panel; the criterion is then applied to these data by a computerized algorithm. Examples of the different types of abstraction forms that are used for the option A and option B methods for applying a criterion are given in Figures 5.3 and 5.4.

Option A (see Figure 5.3), requires less startup work, but training for data collection is more difficult, and more sophisticated abstractors are needed. Inevitably, because more judgment is required of the abstractors, data errors are more common. Option B (see Figure 5.4), requires more startup cost and involves programming the computerized algorithm to apply each review criterion. However, if the system is well built, extensive reviews can be conducted with a high degree of accuracy and speed even when using abstractors with less formal education. (Palmer, Louis, Hsu et al., 1985).

There are other options that combine the features of options A and B. For instance, a structured review procedure can be designed. The abstractor enters the data items selected by the panel, in the order in which

Figure 5.3. Initial evaluation and testing criteria coding: option A (cataract example)

1. Was a complete ophthalmologic exam done *within 6 months after diagnosis?*	Yes___ No___
Exam must include all of the following: appearance of lens, retina/macula, cornea, intraocular pressure.	
If no, did patient refuse exam?	Yes___ No___
2. Was a Snellen visual acuity test done *within 6 months after diagnosis*? (Right and left eyes separately *and* together.)	Yes___ No___
If no, did patient refuse exam?	Yes___ No___
3. Was a glare test done?	Yes___ No___
If yes, was Snellen visual acuity 20/40 or better, *and* patient complained of glare?	Yes___ No___
4. Were any of the following tests done: contrast sensitivity, potential vision, specular photographic microscopy, formal visual fields, fluorescein angiography, external photography, corneal pachymetry, B-scan ultrasonography, electrophysiological tests?	Yes___ No___

Figure 5.4. Initial evaluation and testing criteria coding: option B (cataract example)

Are any of the following documented in the patient record?
Code each exam: 1 = yes, 2 = no, 3 = patient refused

Complete ophthalmological exam

1. __ Appearance of lens — Date __ __/__ __/__ __
2. __ Appearance of macula/retina — Date __ __/__ __/__ __
3. __ Appearance of cornea — Date __ __/__ __/__ __
4. __ Intraocular pressure — Date __ __/__ __/__ __

Snellen visual acuity test

5. __ Snellen visual acuity — Date __ __/__ __/__ __
6a. __ Snellen results, right eye — 20 / __ __
6b. __ Snellen results, left eye — 20 / __ __
6c. __ Snellen results, both eyes — 20 / __ __

Glare testing

7. __ Glare test done?
7a. __ If yes, did patient complain of glare?

Other tests

8a. __ Contrast sensitivity
8b. __ Potential vision
8c. __ Specular photographic microscopy
8d. __ Formal visual fields
8e. __ Fluorescein angiography
8f. __ External photography
8g. __ Corneal pachymetry
8h. __ B-scan ultrasonography
8i. __ Electrophysiological tests

NOTE: This form collects data items that are entered into a computer data base; a computerized algorithm then applies the criteria to the data items, assigning a criterion status to each.

they are most easily found in the data source. A computer program is written to sort these data and print them in a uniform case-abstract format. Clinician reviewers review these easily read case abstracts and apply the criterion, guided by a diagram of the criterion algorithm.

Step 12. Pilot test and revise criteria, forms, and procedures. The performance measure undergoes a process of review and testing. The committee examines the content of the abstraction forms and procedure manuals for face and content validity and determines whether they include all of the data needed to evaluate criteria conformance. To ensure that the terminology and data formats of the forms are consistent with those found in the records, the staff tests the draft instruments by abstracting several patient records.

In a multiorganizational review, care is taken to determine whether abstraction forms and analysis procedures accommodate the broad variety of data configurations found in patient records at different sites. It is

important to identify whether terminology incorporated in the forms or manual is ambiguous or confusing in the special context of any of the review sites. In a pilot test, approximately 30 to 50 patient records are abstracted to provide an adequate representation of the different types of data that may be encountered. A larger number of records may be necessary if there are many different performance sites, because the larger number of test records reveals to a fuller extent the variations in conventions for documenting care that the abstraction forms and coding rules must accommodate.

This pilot test of the performance measure is conducted by the type of personnel who will conduct a large-scale review. Intra-rater reliability is determined when a single, experienced abstractor applies the abstraction forms for a second time to the same case and the two results are compared. Commonly, re-review by the same abstractor is conducted for a 5-percent sample of cases. If the two versions disagree, an explanation for the inconsistency is sought in order to clarify the data abstraction rules and prevent confusion in the future. Inter-rater reliability involves two abstractors of equal skill, each abstracting the same case once and comparing their results. A 5-percent sample of cases is also drawn to examine inter-rater reliability. Initially, a large sample of cases is drawn to pilot test the data abstraction forms and procedures. Both intra-rater and inter-rater reliability testing can help determine how much variation in the data collection is due to the abstraction instruments themselves. In the project to Develop and Evaluate Methods for Promoting Ambulatory Care Quality (DEMPAQ), 100 percent re-review was conducted for the first 100 cases abstracted, and then an ongoing review of 5 percent of randomly selected cases ensured continuing data quality. The 5-percent sample of cases has become the convention for monitoring both accuracy of abstraction and input, as well as continued appropriateness of the forms themselves. (See Appendix C for further details of the DEMPAQ Project.)

The data abstracted during the two types of reliability tests are compared for agreement by using standard statistical packages. These packages calculate the percentage of agreement for each data item reviewed and print out lists of those items that disagree. By identifying the items with high rates of disagreement, the panel staff identify data elements that decrease the reliability of the instrument. They then try to improve agreement (and therefore reliability) through more explicit decision rules in the instruction manual, better training, and/or improved abstraction form design. Re-review and revision of the performance measure take place until the agreement rate reaches an acceptable level, e.g., 95 percent. It is important to note that although 95 percent is often the rate of agreement desired for reliability tests, the acceptable rate may be determined by the requirements of each study. Agreement rate is not a statistical test, so there is no predetermined "right" sample size for the number of cases reviewed. The level of agreement chosen as acceptable depends

upon a tradeoff between the level of agreement desired and the funds available to achieve it.

Once the reliability of the performance measure is established, a process of evaluating its validity is undertaken. A panel of objective clinicians examines the criteria, coding instructions, and abstraction form to determine if they are likely to achieve the purpose for which the performance measure was designed. Clinician reviewers assist in determining the specificity of the measure as a test for guideline conformance by conducting structured implicit review of a sample of the patient records in which the medical review criteria were "not met." They judge whether "not met" cases truly do not conform with the guideline.

Ideally, sensitivity of the guideline-derived criteria should also be tested by conducting structured peer review of cases that met the review criteria. The clinician reviewers are guided to consider these issues:

1. *Are the review criteria appropriate for the performance measure? If not, the reviewers suggest more appropriate criteria so that outdated, inappropriate, or ambiguous terminology may be replaced with more useful terms.*

2. *Are the review criteria applicable to the case? If not, the reviewers give a reason why they do not apply, so that additional exclusions or improved instructions may be incorporated in the abstractor's manual.*

3. *Was the finding of "not met" justified in this case? If not, the reviewers give reasons so that, where appropriate, additional acceptable alternatives or abstracting rules may be formulated.*

An example of detailed instructions for conducting validity reviews is provided in Appendix B.

If the reliability or validity testing reveals that the performance measure is subject to substantial measurement error, each component of the measure is examined for areas of potential correction. The considerations used for revising the review criteria, data sample, and data specifications and procedures are outlined below.

Although the review criteria have had several checks for face validity, invalidity is often identified by the structured review of records as described above. The expert panel examines the suggestions provided by the validity reviewers to determine whether they should be accepted and used to refine the performance measure.

If the number of errors is large because the data sample contains cases with a great variety of characteristics, such as comorbid conditions, the

panel tries to define the data sample more narrowly. Exclusions, such as making certain patients ineligible for the specific criteria set, are added to the definition of the denominator.

To refine data procedures, the panel and staff focus on making the list of acceptable alternatives and synonyms for required data items more comprehensive, and improving the instructions for identifying the appropriate data. If it becomes obvious that new treatments or tests have been adopted by clinicians since design of the performance measure, these are added to the abstraction form and instruction manual. The abstraction forms and instructions are examined for clarity of intent and design. It is desirable to have an experienced abstractor, who has not contributed to development of the particular instrument, use and critique it.

Development of a performance measure is an ongoing process among the expert panel, the technical staff and consultants, and the abstractors. Continuous attention is given to the criteria and data instruments to ensure that they reflect the latest version of the clinical practice guideline and the purpose for which the measure is used. Revisions of the performance measure continue until the amount of measurement error inherent in the instrument is small enough to be acceptable. The amount of error remaining is taken into account when the rates of performance are used.

Implementation Phase

Step 13. Conduct review and assign criteria status. When the committee and staff are satisfied that the performance measure can be applied to the data sample, abstractors use the instrument to collect the data for calculating the performance rates. If paper abstraction forms are used, the results can be hand tallied, or a file of clinical data, or data base is created by entering the data into the computer from the paper forms. Similarly, data from patient or provider questionnaires, administrative data, and other printed sources can be collected on paper forms for entry into the data base.

For data entry, there are a variety of commercially available data base management programs that facilitate the design of the file structure and the data entry screens. To reduce transcription errors, these data entry programs follow the order of the data on the abstracting form and contain checks on the data as they are entered. Data entry checking includes range and format checks and specification of required data.

Using the performance measure developed by the committee, the review analysis is conducted by the quality assurance/quality improvement coordinator (or equivalent individual) or by staff under his/her direction. Training of abstractors at the organizational level usually involves orientation to the purpose of the review, the data sources used, and the review instrument. Such internal reviews can be implemented quickly because abstractors have easy access to and familiarity with their patient records. The close supervision given to staff by the quality

manager means that extensive reliability testing is not required at this level. Criterion status is assessed and coded directly on the abstracting form according to the directions on the form or in the abstracting instructions. This requires abstractors to follow the explicit instructions given for judging criterion conformance.

For multiorganizational performance measures, data may be entered directly from patient records through an interactive computer screen, or electronically transferred from automated patient record systems, computerized claims data bases, and other clinical data bases or registries. Electronic transfer, when possible, produces a data base for performance measurement without the expense of patient record reviews or the potential for errors associated with the abstracting process. In comprehensive integrated medical information systems, such as the Department of Veterans Affairs' Decentralized Hospital Computer Program and the HELP system at the LDS Hospital in Salt Lake City, the routines and data files for quality measurement have been incorporated into the main system.

Training sessions for the abstractors in multiorganizational review are tailored to their level of skill and understanding. Abstractors with different skill levels or experience are taught separately. For those who are less familiar with the data sources and abstraction techniques, it is important to have more examples, more practice time, and time for questions and group discussion.

Step 14. Report review findings. When all the record reviews are complete, the data from the abstracting forms are transcribed onto a table so that they may be summarized. When the rows of the table represent patients and the columns represent criterion status for each criterion in the set, it is easy to calculate the performance rates for small numbers of cases. If the quality assurance/improvement committee routinely uses a simple computer data base program, the data from the abstraction forms are entered into a computer for summarization.

The committee may choose to compute several types of performance rates. For each criterion the rate for exact conformance to the guideline is calculated by dividing the number of cases that meet a criterion by the number of cases eligible for the criterion:

$$\frac{M}{M + AA + NM}$$

where

M = criterion met
AA = acceptable alternative to the criterion
NM = criterion not met

By including the "acceptable alternatives" in the numerator of the rate, the committee permits acceptable alternatives to be considered as conformance to a criterion. Using the same notation, the performance rate is:

$$\frac{M + AA}{M + AA + NM}$$

It is also possible to construct rates of occurrence of each type of acceptable alternative: for instance, if there are five separately numbered acceptable alternatives, the rate for acceptable alternative 1 is computed thus:

$$\frac{AA_1}{M + AA_{1-5} + NM}$$

where

AA_1 = acceptable alternative 1 to the criterion, etc.
NM = criterion not met

For instance, if patient refusal is acceptable alternative 1, observing an unexpectedly high patient refusal rate suggests that providers have had great difficulty in persuading patients to accept a treatment recommended by the guideline.

If a non-conformance rate is desired, the status "not met" is in the numerator, thus:

$$\frac{NM}{M + AA + NM}$$

The denominator for a conformance or nonconformance rate excludes the cases for which the criterion is not applicable or for which there are no data with which to judge criterion compliance. Rates of occurrence of nonreviewable cases, indicating poor documentation of care, can also be calculated:

$$\frac{NR}{M + AA + NM + NR}$$

where

NR = criterion is not reviewable.

Following the data analysis, the quality manager prepares reports of the criteria performance rates for the quality committee. These reports take the form of tables, graphs, and text summaries.

If the committee wishes to summarize performance for a single case for the entire set of guideline-derived criteria, there are several options. A case with nonconformance to any criterion in the set can be scored as not conformant, the possible scores being 0 or 1. This implies an all-or-nothing attitude. The committee can also choose to weight criteria for their importance and use the weighted average of individual criterion scores for a case to sum performance for all criteria for that case. As a third option, the committee can create a rate where the denominator is the number of criteria that apply to a case and the numerator is the number of applicable criteria that were met; this gives a score that is a continuous variable with values ranging from 0 to 1. Rates for cases can then be averaged for the individual clinician, department, or organization that is held primarily responsible, depending on the committee's purpose.

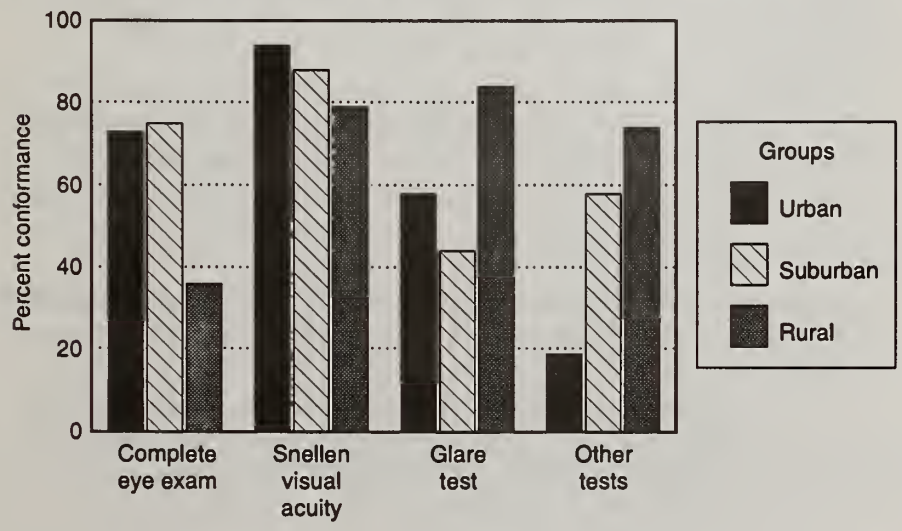
In designing performance rate reports, the committee and staff strive to capture the essential information that users will need to accomplish their purpose. For example, reports for individual clinicians are most helpful if they relate to performance that can be changed by the clinicians themselves. Such strategies for change can include alterations in practice policies, improved documentation design, or enrollment in a continuing education/training course to learn new methods of managing disease or the use of a new instrument.

In reporting rates that are much influenced by system performance (such as timely institution of thrombolytic treatment or timely recall of an ambulatory patient with a biopsy positive for cancer), comparisons between clinicians are only useful to the extent that the clinicians can control these systems, such as physicians who manage their own practices. Such reports are best used to compare departments within organizations or several organizations, and in the spirit of continuous quality improvement, provide performance measurements at repeated intervals of time.

Tables, graphs, and charts are usually used to display large amounts of numerical data. To promote clarity, each table should display a single performance issue. Caution is needed to select a format and measurement scale appropriate to the data, and staff should seek feedback on formats from the group to whom a performance rate will be reported.

Explanatory text and captions accompany the graphs and tables contained in the performance reports (see Figure 5.5). When possible, the text explaining the data is adjacent to the graph or chart. If this is not possible, it should be linked by a clear numbering system. This text includes an explanation of the criterion content, how the criterion was assigned a status, how the performance rate was calculated, the composition of any comparison group

Figure 5.5. Sample report of performance rates for initial exam and testing criteria (cataract example)



whose data are also shown, the derivation of any standards included in the report, and possible interpretations of the data reported.

In multiorganizational reviews, reporting review findings involves three stages. First, criterion status data are analyzed; then performance rates are calculated for each criterion and for combinations of criteria; and finally, reports are issued to feed back the results to clinicians and organizations.

In Steps 8 and 11 above, the derivation of criteria from a practice guideline and the construction of an algorithm for evaluating compliance of individual cases with the criteria is described. The criteria analysis follows the algorithm rules to assign a status of "met," "not met," "acceptable alternative," "not reviewable," or "not applicable" to each of the review criteria.

If the algorithm diagram follows standard flowcharting conventions, a programmer can use it to develop detailed specifications for programming the criterion analysis. The logic applied in calculating a performance rate is revealed by these detailed specifications or by detailed documentation of the criteria analysis program. In this way, the user can determine whether a performance rate is subject to measurement error. For example, the performance rate will be erroneously high if the criterion states that blood pressure and temperature must be taken, but the program logic searches for either blood pressure or *temperature. Conversely, a low guideline conformance rate can be erroneously produced by a program that does not properly identify and remove from the denominator the cases assigned the criterion status "not applicable" or "not reviewable."*

The criteria analysis program produces a computer data file with a status for each criterion for each patient reviewed. The rate calculation program takes these data and aggregates them to produce rates of performance for each criterion for individual clinicians and for groups of clinicians. Embodied in this program are the basic formulas for calculating practice guideline conformance rates.

The rate calculation program can produce rates of performance for individual clinicians. Rates for groups of clinicians are derived by calculating the weighted average of the individual rates of group members. These may be analyzed in various ways, showing rates by organizational affiliation, clinical specialty, geographic area, or other variables of interest.

Prior review projects have identified the following issues as significantly related to the cost and efficiency of programming performance measures:

1. *When large numbers of cases are to be reviewed, the cost of programming is more than offset by the savings achieved by using automated data processing.*

2. *Since computer programming is expensive, time and effort must be allotted for the design and test phases of program development. Careful evaluation at the early stages prevents the wasteful reworking of poorly conceived programs.*

3. *Whether the actual programming is done by internal staff or by consultants, there should be at least one staff member familiar with the performance measure development process who then is responsible for overseeing all the steps in the design, programming, and testing of the systems.*

4. *Training should be tailored to the level of sophistication of personnel, who may range from data abstractors to the clinicians responsible for interpreting performance rate data to others. When possible, professional trainers should be employed, and enough time should be allowed to meet the learning needs of the trainees.*

5. *The project budget should allow for the costs of keeping the software up to date and making revisions as needed.*

6. *Considerable computer memory and speed are required to handle complex, clinically detailed computer programs when large numbers of cases are analyzed. If computer capacity is not sufficient, the computation of performance measurements will be slow.*

These considerations, although obvious, are often neglected in practice.

Computerized systems offered for performance measurement by some software vendors incorporate proprietary techniques. If the company describes what its software does but will not allow inspection of the computer code or provide the algorithm logic, these programs are so-called "black boxes" (Iezzoni, 1991): the user knows only the inputs and outputs for the program, but is blinded to the decision rules employed in producing a "met" or "not met" status for a given criterion. In such instances, the accuracy of the performance rate derived from this criterion cannot be determined, except by a laborious process of sending numerous test cases through the system.

Step 15. Interpret findings, apply standards of quality. The following technical terms are used in this section:

- **Case mix.** A classification of patients into categories reflecting differences in type of illness and/or resource consumption.

- **Confidence interval.** An interval or range based on a random sample, for which there is a given probability (e.g., 95 percent) that the

population mean is contained within that interval. For example, a study may show that a drug lowers the average blood pressure for patients in the study by 4.8 mm Hg, with the 95-percent confidence interval between 2.5 and 7.3 mm Hg. The confidence interval is used in performance measurement to indicate whether an individual rate from a performance review is considered statistically similar to or different from the group average rate, or from a performance rate selected to represent an acceptable level of care.

The committee uses the reports of performance rates to answer the question that prompted the review initially—such questions as "Does the clinic meet the childhood immunization target proposed by the state chapter of the American Academy of Pediatrics?" "Have we maintained the improvement in rates of appropriate treatment of pressure ulcers that was achieved last year?" "Are we doing fewer inappropriate Cesarean sections following our intensive continuing education program?"

Each of these questions implies a standard of quality. The question "Does the clinic meet the target?" uses a prescriptive standard set by an external group respected for its expertise in the subject matter. The questions "Have we maintained. . .?" and "Are we doing fewer. . .?" use a standard set by comparison with prior performance. Standards of quality are applied to a performance rate to decide whether any further analysis or action is necessary. The purpose of the review, who conducts the review, and the strategy for using the review results determine the type of standard of quality that is used. The type of standard, therefore, is determined beforehand and guides the choices made in the planning and development of the performance measure. In the first example, a quality assurance/quality improvement committee in a pediatric clinic constructs its immunization rate with the purpose of determining whether it meets the American Academy of Pediatrics target rate for the State. In the other two examples, a hospital committee compares its current organization's performance with its prior performance in order to determine whether quality improvements are being maintained or are newly achieved.

Comparative standards fit with the strategy of continuous quality improvement, which is driven by the idea of gradually and systematically achieving higher levels of performance. Comparisons to past performance assume that the same performance measure was used to derive the previous rate. If the measure is a new one, a cohort of cases from an earlier time period can be reviewed to establish a measure of prior performance, but it is important for the committee to consider any time trends that would confound the results.

Rates that are compared for different groups of clinicians within the organization are also useful for suggesting future actions. For example, if the rate for pain-free postoperative hours for patients are lower in one unit than in another, the unit with the higher rate will consider adopting the pain control techniques of the unit with the lower rates.

Multiorganizational measurements are useful to describe the clinical performance of groups of clinicians and the systems they work in, and to permit individual clinicians to compare themselves to group norms. The rates of performance from these measurements are interpreted by comparison to rates achieved by other groups, or of the same groups at different times. Cross-comparisons are appropriate only if the reviews were done with the same performance measure, applied in the same way.

Variability observed in rates of performance for specific clinical tasks may be due to patient factors, measurement error, or actual differences in practice between individual clinicians, groups of clinicians, or organizations. Sound interpretation of the rates and using them to improve clinical care require taking into account the measurement error and patient factors that affect measurements made with a particular performance measure. The expert panel attempts to eliminate some of the effects of measurement error and patient factors by carefully specifying the data sample and criteria exclusions. These exclusions remove from the performance measure similar patients who may require slightly different care, because they should be evaluated against different criteria.

Nonetheless, some effect of patient factors may remain on account of differences between the types of patients seen by clinicians who have different types of training or work in different types of organizations. If it were possible to write explicit review criteria so complex that every possible combination of patient circumstances is matched, there would be no need for further exploratory data analysis. In reality, this is not possible. To explore whether differences in performance rates are due to real differences between clinicians, therefore, case-mix adjustments can be made to allow for the contribution of patient factors to the variability shown by the measurement. For example, an assessment of compliance with the acute pain management guideline (Carr, Jacox, Chapman et al., 1992) may show more patients suffering postoperative complications in large referral centers than in community hospitals. If the rates have not been adjusted to reflect the fact that these referral centers generally perform more major surgery, the crude rates will be misleading.

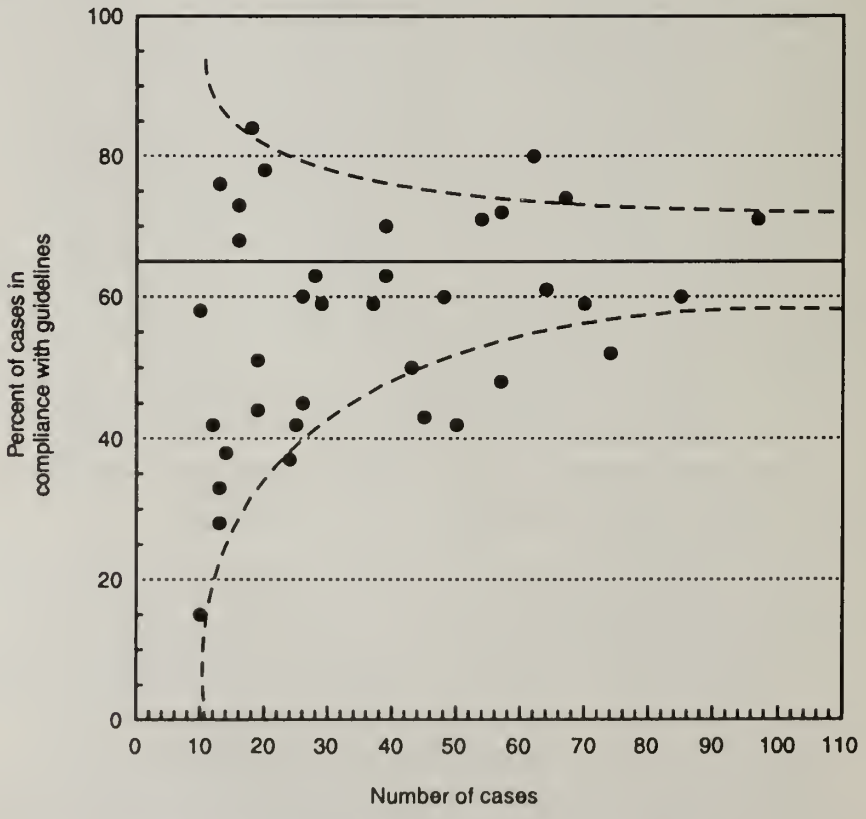
One approach for investigating how various patient, clinician, and organizational factors affect performance is to calculate observed and expected rates of guideline conformance. Risk factors inherent in the patient population (e.g., comorbid conditions) are identified, and overall performance rates for these patient samples are calculated. Then rates for each clinician or clinician group are evaluated on the basis of the proportion of patients who fall into each of the patient risk groups rather than on the basis of comparison with the entire population of patients reviewed. (For further discussion of this issue see Appendix C.)

If the performance of an individual or a group or both is highly variable, the mean performance rate is a poor standard of comparison. Therefore,

if small groups or individual clinicians receive reports of their own performance in comparison with the entire group, the confidence intervals around the individual and the group rates are reported in order to prevent erroneous conclusions. However, a pattern of continued differences between two groups over time may justify further investigation, even if small sample sizes make the confidence intervals too large to show even substantial differences. For an example of a graphic display comparing the performance of different clinician groups (urban, suburban, and rural), see Figure 5.5.

Figure 5.6 is a performance report that displays the performance for a group of providers in relation to the confidence intervals around a prescriptive standard. It is based on an example of a graphic developed by U.S. Quality Algorithms, Inc., and shows hypothetical data. Each data point represents a different provider. The solid line at 65 percent represents a prescriptive standard. The dotted lines show a 95-percent confidence interval around the standard; the confidence interval becomes

Figure 5.6. Report of influenza vaccination of high-risk patients



SOURCE: Adapted from a graphic developed by U.S. Quality Algorithms, Inc.

NOTE: The points mark the average performance rate for individual clinicians. The dashed line marks the 95-percent confidence interval around the prescriptive standard of 65 percent for group performance.

narrower as the number of cases reviewed increases. This approach provides a means to examine the performance of individual clinicians with respect to a standard, but it is not intended primarily for analyzing differences among clinicians. (For an example of the inherent difficulty in comparing individual clinicians with one another, see Appendix C.)

Step 16. Investigate review findings. When performance rates are out of line with the standard that the committee adopts for interpreting the data, a further investigation of the results may be needed. Often, individual patient records are examined by a peer reviewer to determine whether the reported performance rate can be substantiated. If no abstracting errors or reasons why the criteria could not or should not be met are found, the performance rates have indeed identified a quality problem that needs to be addressed. Conversely, if errors are found, the performance measure should not be used again without revising carefully or retraining the data abstractors. Or the committee may ask a quality improvement team to use managerial data sources and special data collection to explore whether, and how, to improve performance by studying systems for delivering care to particular types of patients. Composed of representatives from each department involved in the process, this team explores how performance can be improved.

Although troublesome performance rates have traditionally been investigated and remedied within organizations, there are currently several national programs that can produce multiorganizational comparisons of performance measurements to assist providers in prioritizing quality improvement activities. For example, the Joint Commission on the Accreditation of Healthcare Organizations (JCAHO) has undertaken a program to redesign its standards and accreditation process. This "Agenda for Change" uses nationally uniform performance measures and comparative feedback to enable hospitals to identify areas for quality improvement (JCAHO, 1990). Also, the Health Care Quality Improvement Initiative, introduced in 1992 by the HCFA, requires PROs to develop and disseminate information on patterns of hospital care (Jencks and Wilensky, 1992). Comparisons of hospital rates are intended to stimulate hospital quality improvement efforts aimed at matching the best performance in their State.

Computerized comparisons of performance rates are also used by the Department of Defense in military hospitals and for hospitals of the Department of Veterans Affairs.

One final example of how multiorganizational measurements may be used in a broader context is when State or national specialty societies plan continuing education programs for their members. The project to Develop and Evaluate Methods for Promoting Ambulatory Care Quality (DEMPAQ) was a demonstration project within the PRO program. Aided by a research team, PROs provided comparisons of performance rates and, in collaboration with medical societies, set priorities for continuing

medical education programs. (DEMPAQ is described further in Appendix B, Figure B.1.)

Step 17. Act on review findings. The purpose for which the review was conducted will dictate to a large extent the type of interventions that will follow if quality problems are confirmed. When peer review or investigation by a quality improvement team confirms the existence of a quality problem that was suggested by the performance rates, the quality improvement committee may recommend appropriate action. When responsibility for criterion conformance resides in a specific individual or group, strategies that can be offered by the committee include continuing medical education courses, mentoring, and supervision. When system errors occur, such as failure to recall patients with a suspicious test finding, and responsibility for lack of conformance cannot be attributed to individual clinicians, the quality improvement team is likely to recommend redesigning the system so that it is easier to notice abnormal tests and recall patients.

Step 18. Conduct review again to reevaluate performance. This document primarily addresses use of performance measures for the purpose of quality improvement. Once actions intended to improve quality of care are implemented, the same performance measurements can be repeated to see if the desired improvements have been attained.

Conclusion

The methods described in this chapter can reinforce the use of clinical practice guidelines. Guideline-derived performance measurements can be used to stimulate the activities that are necessary to improve guideline conformance, helping to fulfill the goals of the AHCPR clinical practice guideline program.

References

American Hospital Association. Practice pattern analysis: a tool for continuous improvement of patient care quality. Chicago: American Hospital Association; 1991.

Carr DB, Jacox AJ, Chapman RC, et al. Acute pain management: operative or medical procedures and trauma. Clinical Practice Guideline. AHCPR Pub. No. 92–0032. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services; February 1992.

Diokno A, McCormick K, Colling J, et al. Urinary incontinence in adults. Clinical Practice Guideline. AHCPR Pub. No. 92–0038. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services; March 1992.

Gottlieb LK, Margolis CZ, Schoenbaum SC. Clinical practice guidelines at an HMO: development and implementation in a quality improvement model. QRB 1990;16(2):80–6.

Hadorn DC, McCormick K, Diokno A. An annotated algorithm to clinical guideline development. JAMA 1992;267(24):3311–4.

Iezzoni LI. "Black box" medical information systems: a technology needing assessment. JAMA 1991;265:3006–7.

Jacobs CM, Christoffel TH, Dixon N. Measuring the quality of patient care: the rationale for outcome audit. Cambridge, MA: Ballinger Publishing Co.; 1976.

Jencks SF, Wilensky GR. The health care quality improvement initiative: a new approach to quality assurance in medicare. JAMA 1992;268(7):900–3.

Joint Commission on Accreditation of Healthcare Organizations (JCAHO). Primer on indicator development and application: measuring quality in health care. Oakbrook Terrace, IL: Joint Commission on Accreditation of Healthcare Organizations; 1990.

Longo DR, Bohr D, editors. Quantitative methods in quality management: a guide for practitioners. Chicago: American Hospital Publishing, Inc.; 1991. 136 pp.

Margolis CZ. Uses of clinical algorithms. JAMA 1983;249(5):627–32.

Margolis CZ. Proposal for clinical algorithm standards. Med Decis Making 1992;12:149–54.

Miller MC, Knapp RG. Evaluating quality of care: analytic procedures-monitoring techniques. Rockville, MD: Aspen Publishers, Inc.; 1979.

O'Day DM, Adams AJ, Cassem EH, et al. Cataract in adults: management of functional impairment. Clinical Practice Guideline No. 4. AHCPR Pub. No. 93–0542. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; February 1993.

Palmer RH, Louis TA, Hsu LN, et al. A randomized controlled trial of quality assurance in sixteen ambulatory care practices. Med Care 1985;23:751–70.

Rubenstein LV, Kahn KL, Reinisch EJ et al. Changes in quality of care for five diseases measured by implicit review, 1981 to 1986. JAMA 1990;264:1974–9.

Rubin HR, Rogers WH, Kahn KL, et al. Watching the doctor-watchers: how well do peer review organization methods detect hospital care quality problems? JAMA 1992;267(17):2349–54.

Spath PL. Hospital quality measurement: a story of failure and success. Topics in Health Record Management 1990;10(3):1–22.

Spath PL. Patient care evaluation II. Chicago: American College of Surgeons; 1992a.

Spath PL, editor. Quality management in ambulatory care. Chicago: American Hospital Publishing, Inc.; 1992b.

6. Checklist for Developing Guideline-Derived Evaluation Instruments[1]

The preceding chapters have provided the conceptual and historical framework for the planning, development, and implementation of guideline-derived evaluation tools. In Chapter 4 these three phases are further divided into 18 steps (see Table 4.3), and in Chapter 5 each step is discussed in detail.

Quality improvement (or performance review) committees and their staffs will wish to read these narrative chapters with care. They are important for understanding and completing the tasks required to produce medical review criteria consistent with the guideline, as well as for producing valid and reliable performance measures.

The Criteria and Performance Measure Checklist that follows provides the committees with a means of checking the completeness of their work and documenting their decisions. The checklist is used in a stepwise process as the decisions in performance measure planning, development, and implementation are made. It is not intended to be completed all at once, *nor is it meant as a substitute for a complete and careful study of the main body of this document.*

[1]*Authors: R. Heather Palmer, M.B., B.Ch., S.M.; Naomi J. Banks, M.B.A., M.Ed.; and Patrice Spath, A.R.T., B.A.*

Criteria and Performance Measure Checklist

Planning Phase

Step 1. Clarify the purpose of the performance measurement **see p. 34**

What is the purpose of this performance review (i.e., what do we hope to find out that we don't already know about the performance of health care clinicians or providers)?

_____
_____
_____
_____
_____

Step 2. Identify a relevant clinical practice guideline **see p. 34**

1. What clinical practice guidelines have been developed for the topic we are reviewing?

_____
_____
_____
_____

2. On which clinical practice guidelines will this performance measure be based?

Title:

_____

Source(s):

_____

Title:

_____

Source(s):

_____

Title:

_____

Source(s):

_____

Step 3. Identify populations covered by the guideline see p. 35

1. What patient populations are addressed by these clinical practice guidelines? Consider age category, gender, principal and secondary diagnoses, principal and secondary procedures, level of care setting (e.g., inpatient, outpatient), and other characteristics, and specify as appropriate.

   a. _____
   b. _____
   c. _____
   d. _____
   e. _____
   f. _____

2. Which patient population identified above will be sampled for the performance measure? (Note that the sample will be further defined in Step 6 below.)

   _____
   _____

Step 4. Identify guideline recommendations and draft the medical review criteria see p. 35

1. What are the major sections of the clinical practice guideline?

   a. _____
   b. _____
   c. _____
   d. _____
   e. _____
   f. _____

2. Which of these major sections will be covered in this performance review (if not all)?

   _____
   _____
   _____

3. Detail the clinical practice guideline recommendations in the section that has been chosen for use in the performance measure, the exclusions (if any), and the acceptable alternatives (if any) specified in the guideline. Note that precise definitions of the criteria will be developed in Step 8. (Attach additional page if needed.)

Clinical practice guideline recommendations	Exclusions	Acceptable alternatives

Complete planning phase, review resources needed for development

1. What health care professionals should be represented on the performance review committee or panel?

   1. _____ 6. _____
   2. _____ 7. _____
   3. _____ 8. _____
   4. _____ 9. _____
   5. _____ 10. _____

   Are they represented? ____ Yes ____ No

2. List the technical support staff available for the review:

   1. _____ 4. _____
   2. _____ 5. _____
   3. _____ 6. _____

   Are they sufficient for the design and implementation of the performance measure?

   ____ Yes ____ No

3. What data gathering resources are available for the review within existing budgets?

   1. _____ 4. _____
   2. _____ 5. _____
   3. _____ 6. _____

   Can the review be accomplished with the resources available?

   ____ Yes ____ No

4. What funds are available for the review *over and above* the usual budget?

   _____

5. Who will confirm that these funds can be obtained and spent?

   _____

Development Phase

Step 5. Identify clinicians and sites of care **see p. 38**

1. Which clinicians, providers, and staff will be evaluated (as a group, not as individuals) in this performance measure? List all applicable: title or professional discipline, qualification, or organization.

   _____
   _____
   _____
   _____
   _____
   _____

2. What settings for health care delivery will be evaluated by this performance measure? Check all that apply, and add specifics where appropriate:

   ____ Acute inpatient, specify:
   _____

   ____ Outpatient, specify:
   _____

   ____ Ambulatory/clinic (e.g., freestanding surgery center), specify:
   _____

   ____ Nonacute inpatient, specify:
   _____

   ____ Long-term care, residential care, nursing home, specify:
   _____

   ____ Home care, specify:
   _____

   ____ Other (e.g., psychiatric facility, rehabilitation center), specify:
   _____

3. What is the scope of this performance review for which the measure will be used?

   _____ Single organization _____ Many organizations

Step 6. Define case sample and case sampling period see p. 39

1. Where appropriate, identify the patient characteristics that will be used to select the patient sample to be reviewed. Indicate the data sources to be used to identify patients with these characteristics.

Patient characteristic	Data source
Age category:	
Gender:	
Principal diagnosis:	
Secondary diagnoses:	
Principal procedure:	
Secondary procedures:	
Level of care:	
Other:	
Other:	

2. Specify the time period during which the health care services to be reviewed were delivered (i.e., the beginning and ending dates of the time period in which to look for the event or events selected above):
   From _____ to_____

3. Describe the sampling plan:

   _____
   _____
   _____
   _____

Step 7. Identify data source see p. 42

What data sources (e.g., patient records, claims or billing files, logs, surveys) will be used to measure conformance to guideline recommendations?

_____
_____

Step 8. Write medical review criteria, specifying acceptable alternatives and time window see p. 44

1. Complete the criteria development worksheet (see Table 5.2). (Attach additional page if necessary.)

Data elements	Acceptable alternatives	Data sources	Abstracting instructions/ explanatory notes

2. Record the criteria for measuring performance that were specified on the criteria development worksheet. If you have developed reliable criteria, it should be possible to answer yes to each of the following questions. (Attach additional page if necessary.)

Criterion	Does it reflect the intent of the guideline?	Can a yes-no decision about criterion compliance be made?	Can information necessary to make a yes-no decision be found in the data source(s)?	Have acceptable alternatives and exclusions been clearly identified?	Is a timeframe identified for observation of criterion compliance?

Step 9. Specify data items and data rules see p. 50

1. Record the criteria for measuring performance on the worksheet below. If the criteria are objective, it should be possible to answer yes to each of the questions shown. (Attach additional page if necessary.)

Criterion	Are all data elements clearly defined? Are abbreviations and synonyms provided? Are numeric values included where appropriate?	Is data source identified?	Are instructions and decision rules provided?

Step 10. Draft data collection forms and procedures see p. 53

1. After drafting the data collection instrument(s), ask each of these questions about the tool you will be using for gathering information. It should be possible to answer yes to each of these questions.

   a. Is the instrument organized to permit ease of data collection? Does it follow the organization of the data source(s), or does it have some other logical basis?

   ____ Yes ____ No

   b. Is the manner for recording the requested data elements clearly indicated (e.g., units of measurement, narrative, yes-no)?

   ____ Yes ____ No

   c. When data are precoded, are codes provided on the instrument?

   ____ Yes ____ No

   d. If categories for grouped data are shown on the instrument, are they exhaustive and mutually exclusive (e.g., age = ≤20, 21–25, ≥26)?

   ____ Yes ____ No

   e. Does the visual appearance/layout of the instrument promote understanding and accuracy of data abstraction?

   ____ Yes ____ No

   f. If the data will later be entered into a computer data base, does the visual appearance/layout of the instrument promote ease of data entry?

   ____ Yes ____ No

   g. Are there data abstraction instructions that include all data definitions, decision rules for identifying and abstracting required data, and the method of transferring data into a computer data base for analysis (if relevant)?

   ____ Yes ____ No

   h. If the data will be entered into a computer data base for analysis, does the design of the system facilitate analysis of both criteria status and criteria performance rates?

   ____ Yes ____ No

2. Before pilot testing the performance measure:

   a. Were the following reviewed by the performance review committee or panel and found to have face and content validity:

      (1) Medical review criteria ____ Yes ____ No
      (2) Completed criteria worksheets ____ Yes ____ No
      (3) Data collection instrument(s) ____ Yes ____ No
      (4) Data abstraction instructions ____ Yes ____ No

   b. Was the data collection instrument pretested on representative data by the abstractors who will use it?

      ____ Yes ____ No

Step 11. Devise analysis procedures **see p. 55**

1. What procedure will be used for the data analysis (applying the review criteria to the data items)? Select one:

   ____ Mental application of the criteria by the data abstractor or clinician reviewer.
   ____ Application of the criteria guided by a review algorithm.
   ____ Application of the criteria by computerized algorithm.
   ____ Other, specify:

   _____

2. Does the procedure give unambiguous directions for determining a specific criterion status for each criterion?

   ____ Yes ____ No

3. If a review algorithm flowchart is used, ask each of these questions about the flowchart. It should be possible to answer yes to each of these questions.

   a. Are flowchart symbols correctly used?

      ____ Yes ____ No

   b. Are all questions placed in decision shapes (not on arrows)?

      ____ Yes ____ No

   c. Does each decision shape have two, and only two, outward arrows (one for yes, one for no)?

      ____ Yes ____ No

   d. Is the direction of arrows representing yes and no consistent?

      ____ Yes ____ No

   e. Does each arrow lead to only one next step?

      ____ Yes ____ No

f. Does the algorithm have a clear terminal shape for the entry point?
____ Yes ____ No

g. Does each branch lead eventually to a terminal shape for the end point?
____ Yes ____ No

Step 12. Pilot test and revise criteria, forms, and procedures **see p. 57**

1. If the data collection and analysis instruments were pretested, did the pilot test include an analysis of the following (where applicable)?

a. Revision of the data collection instrument based on the pilot test findings?
____ Yes ____ No

b. Revision of the data collection instructions based on the pilot test findings?
____ Yes ____ No

c. Revision of the algorithm flowchart based on the pilot test findings?
____ Yes ____ No

For multiorganizational performance measure only:

d. Inter-rater reliability of data abstraction?
____ Yes ____ No

e. Intra-rater reliability of data abstraction?
____ Yes ____ No

f. Inter-rater reliability of data entry into computer?
____ Yes ____ No

g. Intra-rater reliability of data entry into computer?
____ Yes ____ No

h. Validity review of criteria labeled "not met" by clinician reviewer?
____ Yes ____ No

i. Tests of computer analysis programs for criteria status and performance rates?
____ Yes ____ No

j. Revision of computer analysis programs based on pilot test findings?
____ Yes ____ No

Implementation Phase

The steps of the implementation phase are influenced by the organization's purpose for measuring performance and by the review findings. For this reason, only common implementation principles are addressed in this checklist. Additional issues pertaining to individual organization must also be considered.

Step 13. Conduct review and assign criteria status **see p. 60**

1. Have all events or persons that were to be included in the denominator of the performance measure been identified and reviewed?

   _____ Yes _____ No

2. Have all required data fields on the data collection instrument been completed?

   _____ Yes _____ No

3a. If data were obtained by data abstractors, did the abstractors follow explicit instructions given for judging criterion compliance?

   _____ Yes _____ No

3b. If data were entered into a computerized data base, did the edit/validation checks confirm the accuracy of the data?

   _____ Yes _____ No

4. If data were obtained through electronic data transfer, did the edit/validation checks confirm the accuracy of the data?

   _____ Yes _____ No

Step 14. Report review findings **see p. 61**

1. Which performance rates will assist the quality committee in meeting its purpose for measuring performance? (See Step 1.)

   a. _____
   b. _____
   c. _____
   d. _____
   e. _____
   f. _____
   g. _____

2. Has the accuracy of the performance rate calculations been validated?

   _____ Yes _____ No

3. After drafting the report of the performance measurement results, ask each of these questions about the report. It should be possible to answer yes to each of these questions.

   a. Is the population clearly described?

   ____ Yes ____ No

   b. Are the performance measures stratified by the appropriate variables?

   ____ Yes ____ No

   c. Are tables or graphs of data accompanied by explanatory text or captions?

   ____ Yes ____ No

   d. Are the sampling methods, confidence limits, and results of statistical tests of data validity and reliability (as appropriate) clearly described?

   ____ Yes ____ No

   e. Are the limitations of the performance measure (e.g., sample bias, case mix, data collection difficulties, measurement error) clearly identified?

   ____ Yes ____ No

Step 15. Interpret findings, apply standards of quality **see p. 65**

1a. Have representatives of all clinicians, providers, or staff who were evaluated by this performance measure been involved in the analysis and interpretation of the findings?

____ Yes ____ No

1b. If some relevant clinicians, providers, or staff are not involved in the analysis and interpretation of the following, what provisions will be made for obtaining their input before Step 15?

_____

_____

_____

_____

2. Has the performance review committee considered the influence of patient factors and measurement error on the variability of performance rates?

____ Yes ____ No

3. What actions are being taken to reduce the impact of patient factors and measurement error on performance rates?

_____

_____

_____

_____

Step 16. Investigate review findings see p. 69

1. If performance rates are not in line with the standard of quality defined by the committee, several different courses of action might be appropriate. Check all that might be considered by the committee:
   - ____ Examine individual patient records by peer review to substantiate the reported performance rate.
   - ____ Form a quality improvement team to explore the relevant processes of patient care.
   - ____ Share comparative performance data with clinicians, providers, or staff who were evaluated by this performance measure to stimulate further analysis and investigation.
   - ____ Other appropriate action as warranted by the performance measurement findings.

Step 17. Act on review findings see p. 70

After completing the analysis and investigation phases, action may be taken if performance changes are required. Ask each of the following questions about the actions. It should be possible to answer yes to each of these questions.

1. Are actions taken on review findings only after careful investigation and involvement of clinicians, providers, and staff who were evaluated by the performance measure?

   ____ Yes ____ No

2. Have the clinicians, providers, and staff who were evaluated by the performance measures been instrumental in designing the actions?

   ____ Yes ____ No

3. Are the actions consistent with the purpose for measuring performance?

   ____ Yes ____ No

4. Do the actions befit the significance of the findings (e.g., the impact of poor-quality performance on patient care outcomes)?

   ____ Yes ____ No

5. Are the actions likely to achieve desired performance changes?

   _____ Yes _____ No

6. Will it be possible to measure the effect of the action through performance improvements?

   _____ Yes _____ No

Step 18. Conduct review again to reevaluate performance **see p. 70**

Appendix B. Validity Review of Performance Measures[1]

Purpose of Validity Review

A structured process for evaluating the validity of a performance measure is essential if the rates are to be relied on for large-scale programs to improve clinical care. In this process a panel of objective clinicians who were not involved in the development of the performance measure examine the criteria, coding instructions, and coding form. Their job is to determine the degree to which a sample of patient records containing criteria that have been "flagged" for nonconformance are truly nonconformant. (A flag is a mark or a code indicating that the review finds a criterion is "not met.") Modifying the performance measure according to the recommendations of the panel provides an important means of reducing measurement error by eliminating obvious sources of false positive findings.

This appendix describes the procedures followed and the materials used to validate the performance measure developed for the project to Develop and Evaluate Methods for Promoting Ambulatory Care Quality (DEMPAQ), a research project contracted by the Health Care Financing Administration. DEMPAQ is described in Figure B.1 (Lawthers, Palmer, Banks et al., 1995; Lawthers, Palmer, Edwards et al., 1993; Palmer, Clark, Lawthers et al., 1994). The DEMPAQ review is conducted by directly entering data from patient ambulatory records into a computer system that displays the abstracting instructions on the monitor.

Methods

Training the Panel

Validity reviewers are provided with a manual of procedures describing the purpose of the performance review and the panel's role in critiquing the performance measure. They are given a two-page review worksheet for each flag that is to be reconsidered; page 1 (Figure B.2) poses the review questions, and page 2 (Figure B.3) provides details of the review criterion and the abstraction process that produces this flag. Reviewers are instructed that the objective of

[1]*Authors: R. Heather Palmer, M.B., B.Ch., S.M.; Ann G. Lawthers, Sc.D.; and Naomi J. Banks, M.B.A., M.Ed.*

examining patient records for validity review is not to make a determination of health care quality for a case but to focus on the ability of the medical review criteria and abstracting instructions to identify guideline conformance. It is emphasized that the review evaluates the criteria, not the clinician.

Figure B.1. Description of DEMPAQ

DEMPAQ
A Project to Develop and Evaluate Methods for Promoting Ambulatory Care Quality

DEMPAQ was a three-year demonstration project to develop tools to review physician office care for the U.S. Government's Peer Review Organization (PRO) program. PROs are organizations of physicians in active practice who contract with the U.S. government to review quality of care in the Medicare program, which serves persons over 65 years of age.

The project emphasized education and feedback to physicians and concentrated on the quality of care rendered in office settings. The study population was a random sample of primary care physicians with an office practice who treat Medicare patients in the states of Maryland, Iowa, and Alabama.

DEMPAQ developed two review methods: a review of medical records and an instrument for creating profiles from a national data base of medical care billing information. The following is a summary of the methods used for review of medical records.

The record review focused on functions the physician typically performs in patient encounters: ordering tests, prescribing drugs, performing procedures, making new diagnoses, following up on abnormalities, and offering preventive or screening care. Rates of satisfactory performance are averaged across patients to yield a profile for the physician. The rates are generated by abstracting data from medical records and applying to them computerized algorithms, one for each function.

Personal data are reported to providers along with the corresponding rates for their peers. This educational component is critical to the philosophy of the project. The opportunity of this feedback is used to solicit comment from physicians whose performance was reviewed on the usefulness and value of the review.

Critical to the design of the project is that input from physician organizations has been sought in both the design and evaluation phases of the project. The record review criteria have been submitted to liaison members from these organizations for critique and comment before implementing the profiling or record review. Following the feedback phase, the same physician organizations are asked to evaluate the educational value and usefulness of the data.

In addition to developing the record review, DEMPAQ has undertaken extensive evaluation of both the validity and reliability of the review technique. The patient record data are also compared to that contained in the HCFA-1500 to assess whether claims data can be reliably used to assess quality.

Figure B.2. DEMPAQ peer review worksheet, page 1

Peer reviewer: _____
Date: _____
PRO: _____

PATIENT ID: 0014422 FLAG: D2
CODE: 2
FLAG SCORED FOR:
DIAGNOSIS: Ischemic Heart Dis on 11/11/1988
REASON:
New diagnosis not justified by evidence in medical record.
REFERENCE:

The patient care episode shown on the label above was flagged for physician review because it failed to meet the DRS review criteria. On the attached page you will find the review format and the criteria used by the initial PRO reviewer. Please review the medical record for the item of care shown above, and indicate below your own judgment about the item in this case.

CONSIDER: Are the review criteria appropriate for this indicator?
Are the review criteria applicable to the case at hand?
Was the flag justified in this case?

Choose only the one response that most closely fits your opinion. Use the spaces below and the back of *this* page to give your comments.

____ 1. The criterion flag was *justified* given the data in the medical record.

____ 2. The criterion could not be met in this circumstance because of *patient factors*. Describe the patient factors below under "Comments."

____ 3. The criterion could not be met in this circumstance because of *practice factors*. Describe the practice factors below under "Comments."

____ 4. The criterion should not be met in this case due to *extenuating circumstances*. Describe the extenuating circumstance below under "Comments."

____ 5. The flag is not justified because of *abstractor* error. Describe the error under "Comments."

____ 6. The flag is not justified because the abstracting *decision rule is incorrect or incomplete*. Suggest a revised decision rule below.

____ 7. The flag is not justified because the *criterion needs to be revised*. Propose a revised criterion for this item below.

Criterion/Coding rule revisions: _____

_____

Comments: _____

Continue comments on back

Is the documentation in this record adequate to conduct a valid quality of care review? ____

_____

Figure B.3. DEMPAQ peer review worksheet, page 2

Ischemic heart disease FLAG: D2
CODE: 02

Review question displayed on DRS screen

Evidence for a New Diagnosis
Patient ID: Diag.: Visit date:

Indicate if any of the following evidence for a new diagnosis is in any note in the medical record: 1=yes 0=no

Look in entire medical record.

0◀

Clinical criteria displayed in window on DRS screen:

Evidence establishing a new diagnosis:

1. Cardiac chest pain (worse with exercise, relieved by rest) with ST segment changes on ECG

   OR

2. Positive exercise tolerance test

   OR

3. Positive cardiac catheterization

Reviewers are instructed to consider the following broad concerns as they examine very specific questions about the clinical content of each criterion:

1. Are the review criteria appropriate for the performance measure? If not, the reviewers are asked to suggest more appropriate criteria so that outdated, inappropriate, or ambiguous ones may be replaced with criteria that are more useful.

2. Are the review criteria applicable to the case? If not, the reviewers are asked to give a reason why the criteria do not apply, so that the abstractor's instructions for the review process can be improved.

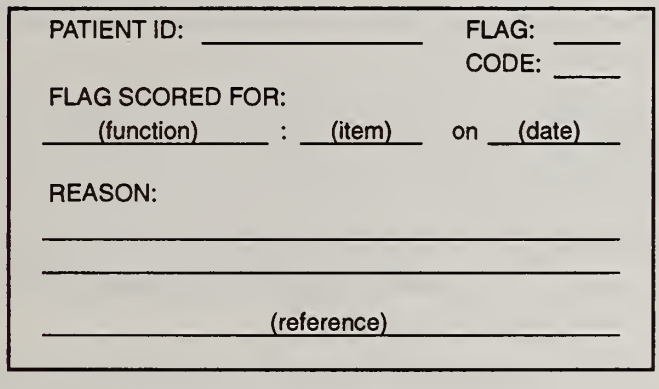
3. Was the flag (i.e., the coding that the case did not conform to the guideline) justified? If not, the reviewers are asked to give reasons

so that, where necessary, additional acceptable alternatives or abstracting rules may be formulated.

Review Worksheet, Page 1

In addition to the instructions in the manual, DEMPAQ reviewers are provided with the patient record and a two-page review worksheet (see Figures B.2 and B.3). At the top of page 1 of the review worksheet is a 2- by 4-inch self-adhesive label (Figure B.4) generated by the computer on the basis of the case reviews stored within it. The label bears information identifying the case and the flagged item to be reviewed. Flags are generated each time a review finds that the practitioner did not meet a criterion for one of the indicators associated with the item.

Figure B.4. Computer-generated label



The entries on the label have the following meanings:

- *Flag.* The code for the criterion (e.g., D1, T3).
- *Code.* The code of the clinical item (e.g., hypertension [HTN], diabetes mellitus [DM], hematocrit [HCT], electrocardiogram [ECG]).
- *Function.* One of the 6 physician activities assessed by DEMPAQ:
  - prescribing drugs
  - making new diagnoses
  - ordering tests
  - performing procedures
  - following up on abnormalities
  - offering preventive or screening care
- *Item.* A particular clinical service related to this activity; e.g., for ordering tests, one of the clinical items is a hemoglobin test.
- *Date.* The date of the visit note in which the item was observed by abstractor.

- *Reason.* A brief remark describing the reason for the flag (e.g., "Actions not taken to monitor diagnosis appropriately," "No appropriate followup of abnormal test results").

- *Reference.* Additional information for some flags that assists in reviewing the flagged item.

Below the label is a list of possible evaluations of the flag. After reviewing the patient record with the assistance of the information on page 2 of the worksheet, reviewers are asked to select the one response that fits most closely their evaluation of the flagged item and to explain their response with comments. The same choice of responses is offered for all flags:

1. The criterion flag was *justified*; the criterion is appropriate.

2. The criterion could not be met in this circumstance because of *patient factors*. The reviewer is asked to describe the patient factors under "Comments."

3. The criterion could not be met in this circumstance because of *practice factors*. The reviewer is asked to describe the practice factors under "Comments."

4. The criterion should not be met in this case due to *extenuating circumstances*. The reviewer is asked to describe the extenuating circumstance under "Comments."

5. The flag is not justified because of *abstractor error.* The reviewer is asked to describe the error under "Comments."

6. The flag is not justified because the abstracting *decision rule is incorrect*. The reviewer is asked to suggest a revised decision rule.

7. The flag is not justified because the *criterion needs to be revised.* The reviewer is asked to propose a revised criterion.

Examples of how to select the proper response are shown below.

Review Worksheet, Page 2

Page 2 is specific to the flagged clinical item being reviewed. There are many criteria-item combinations (e.g., *monitoring* a diagnosis of *diabetes*, *followup* of abnormal *creatinine*). Page 2 of the worksheet conforms to the case being reviewed, as revealed by the information on the label on page 1.

The format used by the abstractor to answer the questions that produced the flag is shown on the computer monitor and reproduced at the top of page 2. This is the generic format of the question seen by the initial reviewer when evaluating the criterion. Below the illustration of the format on the computer

monitor, the worksheet shows the text of the criterion used for review of that specific clinical item. By seeing the actual review questions that appear on the data entry screen, the list of possible responses, and the clinical criteria unique to that item, the validity reviewer attempts to reconstruct the rationale for the original review that resulted in a flag. This exercise assists the reviewer in assessing the validity of the criterion.

Review Coordinator

A review coordinator, who is an experienced reviewer, assists the review panelists by preparing the worksheet and the patient record for the review and is present during the review to answer any questions that the clinician reviewer might have. The coordinator attaches the computer-generated label to page 1 and selects the appropriate second page for the worksheet to reflect the criterion-item combination specified on the label. Chart preparation involves locating the visit date in the record and the clinical item identified on the worksheet label. Relevant clinical information is highlighted in the patient record to reduce the reviewer's time searching the record. The review coordinator is familiar with the abstracting rules used by the original peer review organization reviewer and thus can clarify any questions about the way the original review was conducted. This strategy of using a review coordinator to facilitate the work of the validity reviewer has made the review efficient, allowing the clinician to concentrate on evaluating the criteria.

Examples of Validity Review Evaluations

Figure B.5 is an excerpt from the reviewer instruction manual, showing examples of situations that would result in one of the criteria judgments that may be selected from the list on page 1 of the worksheet.

Using the Validity Review Evaluations

The panel staff summarize the validity reviewers' evaluations and comments. The panelists examine these summaries to determine whether any part of the performance measure should be modified. Suggestions for improving data abstraction forms and instructions are incorporated into the measure.

Suggestions regarding the criteria are handled in several ways. If the validity reviewers identify inconsistencies between the criteria and the guideline on which the performance measure is based, then the criteria may be revised accordingly. However, if the suggestions for changing criteria are contrary to guideline recommendations, and clinician convenience or habit is the reason given for rejecting the guideline, the criteria should *not* be modified. The panel notes the discrepancy and the number of reviewers who recorded

Figure B.5. Excerpt from the DEMPAQ reviewer instruction manual

1. The criterion flag was *justified;* the criterion is appropriate.

   Example The action required by the criteria is appropriate for the patient in this case. You have searched the specified timeframe in the patient record, and do not find the action. The review was apparently conducted accurately, and you do not disagree with the criterion.

2. The criterion could not be met in this circumstance because of *patient factors.*

   Example A The action required is appropriate, but is not found in the patient record. The record states that the patient was in Florida for two months during which the action was to have been performed.

   Example B The action required is appropriate, but is not found in the patient record. The record states that the patient was diagnosed with terminal lung cancer, making a mammogram unnecessary.

3. The criterion could not be met in this circumstance because of *practice factors.*

   Example The action required is appropriate, but is not found in the patient record. There is evidence in the record that the practice was in the process of converting to a new computerized practice management system, and the procedures for patient recall were disrupted during this timeframe.

4. The criterion should not be met in this case due to *extenuating circumstances.*

   Example The action required is appropriate, but is not found in the patient record. There is a note that the patient is undergoing psychotherapy and has initiated legal proceedings against her husband following an episode of domestic violence. In this case, the patient's psychosocial problems supersede the strict timeframe of the criterion.

5. The flag is not justified because of *abstractor error.*

   Example A The action required is appropriate, but is not found in the patient record. There is a note that the patient refused an exam. This is a coding error. If the reviewer had coded "8" for patient refusal, the item would be scored as an "acceptable alternative," and not flagged.

   Example B You found evidence in the record that the action was done. The initial reviewer appears to have seen that visit note because other data from the visit have been used in the review.

6. The flag is not justified because the abstracting *decision rule is incorrect.*

   Example The action specified for monitoring diabetes is exam of the fundi. This is appropriate, and the patient record shows that "HEENT" were examined. The case is flagged because the abstracting rule is that for diabetes monitoring, a fundus exam should be stated explicitly. You believe that HEENT is sufficient evidence of a fundus exam, and do not think the case should be flagged.

7. The flag is not justified because the *criterion needs to be revised.*

   Example You disagree with the clinical content of the criterion as stated on the bottom of page 2 of the worksheet. Give your rationale and suggest a more appropriate criterion.

criticism of the criterion. Interpretation of the performance rates for these criteria is influenced by the strength of disagreement with them.

For example, the American Board of Family Practitioners guideline for diabetes management recommends quarterly HgbA1C monitoring. Even though a large proportion of validity reviewers might suggest that a criterion to this effect should be changed to permit serum glucose testing instead, changing the criterion would be inappropriate, because it is based on a published guideline. However, since the reviewers as a group feel so strongly about this point, the discussions of performance rates by the users of the measure include a consideration of the differences between recommended and actual practice.

References

Lawthers AG, Palmer RH, Banks N, et al. Designing and using measures of quality based on physician office records. J. Ambulatory Care Manage 1995;18(1):56–72.

Lawthers AG, Palmer RH, Edwards JE, et al. Developing and evaluating performance measures for ambulatory care quality: a preliminary report of the DEMPAQ project. The Joint Commission Journal on Quality Improvement 1993;19:552–65.

Palmer RH, Clark L, Lawthers A, et al. DEMPAQ: a project to develop and evaluate methods to promote ambulatory care quality. Final Report, vols. 1, 2, 3. Grant under contract with the Health Care Financing Administration, contract 500–89–0624, Easton, MD: Delmarva Foundation for Medical Care, Inc.; 1994.

Appendix C. Statistical Issues for Rate-Based Measurement[1]

Introduction

This appendix is for readers who have some background in statistics and who are interested in quantitative approaches to the measurement of guideline conformance. Three issues are introduced: (1) how to choose a random sample; (2) how to adjust or standardize rates in order to make comparisons "fair"; and (3) how to compare rates while appropriately taking variability into account. Examples are constructed and assumptions are made to illustrate these concepts, but implementing them will require adaptation to the characteristics of each specific data set. So, for example, while rate standardization is illustrated by adjusting rates of appropriate coronary artery bypass graft (CABG) procedures for anesthetic risk and by adjusting rates of conformance to an asthma guideline for severity of an asthmatic attack, it is not suggested that these are the only two factors that need to be considered, or that these two factors would be appropriate for other rates. The purpose of such illustrations is to open up these ideas for discussion and point out that the required analytic techniques exist.

Sampling Issues

Performance rates are calculated from only a sample of patients, not from all possible patients. Hence a strategy is needed for drawing the sample, and a mechanism is needed for carrying out the strategy. Finding the optimal strategy can be difficult, depending on the goal or goals of the study. The mechanism is much simpler: at some stage it is simply an issue of choosing a random sample. Two sampling strategies—pure random sampling and stratified random sampling—are considered here, and their advantages are contrasted. The mechanism for choosing a random sample is described first.

Given a pool of patients, the goal of random sampling is to choose a certain number of patients in such a way that every patient in the pool has the same probability of being chosen:

[1]*Author: E. John Orav, Ph.D.*

> If there are a total of N patients in the pool and K of them are chosen, then each patient could have been chosen with probability K/N.

Following this principle protects against biased samples that could lead systematically to false conclusions. For example, choosing the first patient who appears each day could lead to systematically different results than choosing the last patient who appears each day. Random sampling ensures that the sample is, on average, representative of the patient pool. Even with random sampling, in any given instance bad luck may deliver a sample with extremely good or bad characteristics. However, this impact is not systematic, and it is accounted for in the measures of variability and confidence intervals discussed next. More details on the advantages and disadvantages of random sampling are given in Longo and Bohr (1991).

There are two possible methods of choosing a random sample, depending on whether the entire patient pool is already identified (e.g., all patients seen last year) or the patients are accruing as the sample is being drawn (e.g., all patients who are being seen this year). If the entire patient pool of N patients is already identified, then their names are listed from 1 to N. A random number from 1 to N is generated, and that number is matched to the list of patients, leading to the first "chosen" patient. (A random number can be generated through any of several computer programs, including most data base and statistical analysis packages.) A second random number is then generated; if it is the same as the first, it is disregarded; if it is new, it identifies the second patient chosen. This process continues until K patients are chosen. The original order in which the patients were listed is irrelevant.

If a random sample of patients as they accrue is desired, then the true size of the patient pool will be unknown, and the problem is usually phrased in terms of the percentage of patients whose records are to be reviewed. For example, assume that records from 10 percent of the patients will be inspected. Every time a patient appropriate for the pool is identified, a random number between 0 and 1 (again using one of the sources discussed above) is generated. If the random number is less than or equal to .1, then the patient is chosen; if the random number is greater than .1, then the patient is not included in the random sample. When the next appropriate patient is identified, a new random number is generated, and the process is repeated. At the end of the experiment, about 10 percent of the patient population will have been chosen to be part of the sample.

A primary issue in calculating performance rates is deciding from which pool of patients to draw the sample. Assume first that a performance rate that characterizes an individual (e.g., a single clinician, a group practice, or a hospital) is to be constructed. The simplest procedure would be to take a random sample of all the patients seen by that individual. The rate would then be calculated as the number of patients who met the criterion of interest (e.g., the number of patients who had an appropriate CABG procedure, or C) divided by the total number of patients in the sample, or N. This estimated rate would, in sta-

tistical terms, be unbiased; in conceptual terms, the estimated rate would tend to be neither systematically higher nor lower than the "true" rate of appropriate CABG procedures. Hence this estimated rate would be a fair representation of the chance that the next patient would receive an appropriate CABG procedure. This is an example of a pure random sample, leading to an estimated rate we will call R = C/N.

As an alternative, the same practitioner could be evaluated by taking a stratified random sample of his or her patients. For simplicity, consider dividing the patient pool into two strata: patients with comorbidities indicating high anesthetic risk, and patients whose lack of comorbidities indicates low anesthetic risk, assuming that restricting CABG procedures to appropriate indications is desirable in patients with high anesthetic risk. Any categories that divide patients into subgroups with internally homogeneous responses can be chosen. It is assumed that the percentage of the individual's patients who are high risk (P1) and the percentage of patients who are low risk (P2) are known. A random sample of N1 high-risk patients and a random sample of N2 low-risk patients are then selected. If C1 of the N1 high-risk patients had appropriate CABG procedures, and C2 of the N2 low-risk patients had appropriate CABG procedures, the overall rate of appropriate CABG procedures can be estimated as:

$$R2 = P1*(C1/N1) + P2*(C2/N2)$$

It is not difficult to show that this rate is also unbiased and shows a fair estimate of whether an appropriate CABG procedure will be done by a clinician. The advantage here is that the rate is unbiased regardless of how N1 and N2 are chosen. Choosing N1 and N2 wisely can reduce the variability of R2 in comparison with the variability of R above. Therefore, R2 will be a better estimator, since it is more likely to be close to the true rate of appropriate CABG procedures. In one special case, the two estimation methods will give essentially similar answers. This occurs when N1 is chosen so that N1/(N1+N2) = P1, and N2 is chosen so that N2/(N1+N2) = P2. However, choosing a larger sample size for the stratum in which the performance rate is closest to 1/2 results in a smaller variance for R2 than for R. The optimal sample sizes depend on the true performance rates in each stratum, which are not known initially. Making reasonable guesses about the true performance rates, and so choosing appropriate sample sizes, can often decrease, if not minimize, the variability of the rate.

Another advantage to using a stratified random sample is that it ensures that an adequate number of patients is reviewed in each of the relevant strata (i.e., N1 and N2 are both large enough) so that reasonably accurate estimates of appropriate CABG procedures in high-risk patients (C1/N1) and in low-risk patients (C2/N2) can be determined. Obtaining accurate estimates may not be possible with pure random sampling, since the percentage of any particular subgroup of patients may be so low that a pure random sample would contain

only a few such patients. A rate based on only a few patients would be so variable as to be worthless.

The major difficulties with stratified samples are practical ones. For example, the relevant strata must be identified in advance, and the patient pool must be divided according to the strata. Such information is not always easily available. The sizes of the samples from the different strata must be based on prior guesses regarding the true rates in those strata. Further, the overall rate must be reconstituted from the individual rates and the percentage of patients in each stratum; this calculation is more difficult conceptually and mathematically than the simple average used after simple random sampling.

A final issue concerning sampling follows naturally from the idea of stratified random sampling: If a large system such as an entire State is being evaluated, should a random sample be drawn from the pool of all patients in the State, or should the patients be divided into smaller strata (e.g., hospitals, or practice groups, or even individual clinicians) before a stratified random sampling is drawn? If a pure random sample is drawn from across all patients in the State, then clinicians or institutions that see many patients will be heavily represented in the sample. This is fine, although it should be remembered that the estimated rate represents the chance that the "next" patient receives inappropriate care. Since the "next" patient is more likely to go to a high-volume clinician or institution, that clinician's or institution's performance rate should weigh more heavily in the patient's potential outcome. Moreover, enough patients can be reviewed so that the variability of the performance rate estimate will be low, and the rate will be pertinent. One of the negative aspects is that after pure random sampling, rates for smaller subunits (e.g., clinicians) cannot be examined because there may be many low-volume clinicians for whom no or very few patients have been sampled.

On the other hand, if stratified random sampling is used, the overall performance rate as a weighted average of the individual performance rates from the individual clinicians can be reconstituted as long as the volume for each individual clinician is known. Moreover, it can be guaranteed that a certain minimal number of patients is reviewed from each clinician. As will be seen from the discussion of confidence intervals, the danger is that many patients from an individual clinician are needed before the variability in the estimated rate drops sufficiently for results to be interpreted with confidence. For example, a statewide rate based on 1,000 patients would provide a very stable and interpretable estimate. However, if the 1,000 patients were sampled stratified as 10 patients from each of 100 clinicians, the estimate for each clinician would be unstable. The confidence bounds for the individual clinician rates would almost certainly overlap a State average rate, sending a message that everything was fine. In fact, there is virtually no statistical power to detect discrepancies among clinicians.

In summary, the sampling scheme should be chosen coupled with the variability considerations described. Stratified random sampling provides a

tool that can improve power and allow closer, more detailed analyses. At the same time, focusing the sampling at too detailed a level without increasing the sample size to provide sufficiently tight confidence bounds can produce misleading negative findings.

Exploring Variability When Interpreting Performance Rates

Conformance with guidelines is measured by constructing clinician profiles (i.e., clinicians' rates of performance on specific tasks). A performance rate, however, can best be interpreted through comparison with an external standard. A fair comparison between clinicians or clinician groups requires allowing for differences due to patient case mix, as described later in this section. This adjustment of the estimate should also be supplemented with confidence limits, allowing for variability due to both sampling error and clinician-related differences, such as differences in the patients served by different clinicians. (To avoid confusion with the concept of standards of quality discussed in relation to performance measures, the terms *adjusted* and *adjustment* are used throughout this section in place of the epidemiological terms *standardized* and *standardization.*)

For example, assume an HMO is being characterized according to the number of acute asthmatic attacks with treatment not conforming to guideline in 1 year. N represents the number of patients seen for an asthmatic attack during the year; C represents the number of patients who had nonconforming treatment. Then a simple summary profile is:

$$\text{Rate of nonconforming treatment} = C/N$$

It is not necessarily useful to compare such a simple summary rate to a standard of performance or to compare it between HMOs, because different HMOs may see different types of patients, for whom receiving the recommended care is more or less critical. For instance, a committee may be less likely to recommend expending resources on quality improvement if nonconformance to treatment guidelines is restricted to mild asthmatic attacks that might well resolve spontaneously. Before making any comparisons, the committee therefore adjusts for severity of the asthmatic attacks. Patients are divided into distinct categories, and HMOs are expected to perform similarly when treating patients within such a category. The case-mix adjustment is made by using either of the following:

- *Direct adjustment*, in which the HMO profile is reconstructed, adjusting the rate to reflect a standardized case mix.
- *Indirect adjustment,* using national rates within each category of patients and calculating the expected rate for the case mix observed for a given HMO. The observed rate is then divided by the expected rate.

While either method could be used in principle, direct adjustment may not be possible if there is only a small number of patients for each clinician. In this case, the clinician performance rate cannot be calculated if no patients for a particular category were reviewed. When small sample sizes are common, it is more practical to pursue indirect adjustment.

In an example of indirect adjustment, patients with acute asthmatic attacks are categorized by the "severity" of the attack (severity determined by peak flow rate): category 1, mild asthma; category 2, moderate asthma; and category 3, severe asthma.

Then,

N1 = the number of patients in category 1, mild asthma;

N2 = the number of patients in category 2, moderate asthma;

N3 = the number of patients in category 3, severe asthma;

and

R1 = proportion of patients with mild asthma whose care does not conform to guideline;

R2 = proportion of patients with moderate asthma whose care does not conform to guideline; and

R3 = proportion of patients with severe asthma whose care does not conform to guideline.

The expected number of nonconforming treatments for the observed HMO is

$$E = (N1 \times R1) + (N2 \times R2) + (N3 \times R3)$$

This forms the indirectly adjusted rate:

$$\begin{aligned} &\text{Observed rate/expected rate} \\ &= (C/N) / (E/N) \\ &= C/E \end{aligned}$$

This indirectly adjusted rate has a simple interpretation. If the rate equals 1, then the HMO is performing as would be expected according to the standard. On the other hand, if the adjusted rate is greater than 1, then the HMO's profile is higher than would be expected. And if the adjusted rate is less than 1, then the HMO's profile is lower than would be expected.

While the adjusted rate is accurate and interpretable as a summary of HMO performance, adjusted for patient case mix, the indirectly adjusted rate

remains only an estimate based on a sample of the HMO's patients. The more patients in the sample, the more likely the estimated rate is close to the true rate for that HMO. However, with a finite sample the estimate can be far from the truth. In providing an HMO with a report of its performance, providing upper and lower limits between which the truth is (probably) included (as well as the estimated rate) would be important. To provide such limits, a confidence interval is created around the estimated rate.

A confidence interval would also have a natural and familiar interpretation. If the confidence limits contain 1, then the HMO could be performing at the standard even though the estimated rate may be different from 1. On the other hand, if the confidence limits do not contain 1, then the HMO is probably performing at a rate above or below (as indicated by the estimated rate) the standard. The confidence limits are constructed according to the methods in Breslow and Day (1987).

It is assumed that the total number of nonconformant treatments in a year has a Poisson distribution (a common distribution describing rare events). Then a 95-percent confidence interval for the indirectly standardized rate is given by the lower and upper bounds, in which

$$\text{lower} = [C/E]*[1 - (1/9C) - (1.96/3C^{1/2})]^3$$

and

$$\text{upper} = [(C+1)/E]*[1 - (1/9\{C+1\}) + (1.96/3\{C+1\}^{1/2})]^3$$

Carrying through the previous example, it is assumed that the adjusted ratio for nonconformant treatments, C/E (C, the number of nonconformant treatments, over E, the expected number of nonconformant treatments), equals 1.25. Consider HMO A, for example, for which the actual number of nonconformant treatments is 60 and the expected number of nonconformant treatments is 48, for a 95-percent confidence interval between .95 and 1.60. With the low sample size of 60 cases, there is not enough evidence to reject the possibility that the HMO is performing within the standard, even though its nonconformant treatment rate is higher than the normal (1.25). Now consider HMO B, whose number of nonconformant treatments is 125 and expected number of nonconformant treatments is 100, which places the 95-percent confidence interval between 1.04 and 1.49.

With the larger sample of 125 cases there is enough evidence to state that HMO B's nonconformant treatment rate, although it is the same rate as HMO A's (1.25), is significantly higher than standard.

One possible problem may remain. The confidence limits that were calculated allow only for sampling variability. Would it be expected or desirable for all HMOs to perform exactly the same? There may be a number of reasons why this might be an unreasonable expectation for any particular HMO:

- The case-mix adjustment may not be perfect.
- There may be regional variations related to varying prevalence of the guideline-related disease or relevant comorbidities.
- There may be variations in access to facilities or third-party coverage that affect patient willingness to cooperate in treatment.
- There may be variability in the application of the review criteria.

There might be a certain amount of natural, even expected, variability among the "true" performance rates of HMOs. The importance of cofactors that might account for this variability can be explored, for instance, by expanding the confidence limits for a particular HMO to allow for this natural variability as well as for the sampling variability considered previously.

To make this allowance, the following hierarchical model has been developed at the Harvard School of Public Health for use in comparing performance among clinicians in the DEMPAQ project (Lawthers, Palmer, Edwards et al., 1993). It is still assumed that the number of nonconformant events for a given clinician follows a Poisson distribution. It also is assumed that the mean of the Poisson distribution is specific to each clinician. Then, among different clinicians, the Poisson means are themselves random variables drawn from an exponential distribution.

Under these assumptions, a standard calculation for mixture distributions shows that, for the clinician for whom a profile is being created, the number of nonconformant events follows a geometric distribution. The mean of the geometric distribution depends on the amount of variability allowed between clinicians.

The amount of additional variability introduced through these assumptions is illustrated as follows:

- If only sampling variability is assumed, then the number of events of guideline-related interest is drawn from a Poisson distribution and will have a variance best estimated as "C."
- If the hierarchical model is assumed, allowing for variability between clinicians as well as sampling variability, then "C" will have a variance best estimated as $C + C^2$.

This additional variability becomes incorporated into wider confidence intervals. Assume that the C nonconformant treatments occurred as follows: C1 among patients with mild asthma attacks, C2 among patients with moderate asthmatic attacks, and C3 among patients with severe attacks, and C1 + C2 + C3 = C. Then, under the hierarchical model the variance of C equals $C + C1^2 + C2^2 + C3^2$. This is in contrast to the simpler assumption of only sampling variability, which says the variance of C equals C.

The variance from the hierarchical model is then used to construct the 95-percent confidence intervals for the indirectly adjusted rate, using an approximation in which the lower and upper limits are as follows:

$$\text{Lower} = (C/E) - [1.96*\{\text{variance}(C)^{1/2}\}/E]$$

$$\text{Upper} = (C/E) + [1.96*\{\text{variance}(C)^{1/2}\}/E]$$

As an illustration of the impact of the additional variability, it is assumed as before that C equals 60 and E equals 48, with C1 equaling 30, C2 equaling 20, and C3 equaling 10. Then the 95-percent confidence interval, with only sampling variability, is [.93, 1.57]. In contrast, the 95-percent confidence interval that allows for clinician-to-clinician variability as well as sampling variability is [.45, 2.05]. It is recommended that these hierarchical intervals be developed further and considered for use in reporting guideline conformance.

Summary and Future Considerations

The accuracy of the confidence intervals described in the previous section is directly tied to the sampling scheme and sample size employed. As seen in the examples, large numbers of patients may be required to construct tight bounds that allow accurate discrimination between "normal" and "abnormal" rates. That is, if a threshold standard were set for further investigation or intervention, large numbers of patients may be needed before we are confident that we can distinguish acceptable from unacceptable rates. If stratified sampling is done, with an intent to analyze rates separately in each stratum, then these large sample sizes must be available in *each* stratum. As a particular, common, example, stratifying by clinician in order to evaluate the performance of each clinician separately may require a very large review effort to ensure meaningful comparisons. Often, attempts to evaluate individual clinicians require such large sample sizes that the studies either become unworkable or are carried out with insufficient power to lead to useful conclusions. It is generally more practical to evaluate clinicians as a group, such as the staff of an HMO or hospital, where large-scale sampling is practical, using stratification to ensure adequate representation in terms of patient case mix.

Although this statistical discussion has been restricted to the construction of confidence intervals, the same issues and remarks apply to decisionmaking and the implementation of intervention programs. The same small sample sizes that create wide confidence bounds and prevent a decision about whether a particular clinician is truly performing according to standards also create power problems when a decision is needed to implement an intervention program. A natural analog to developing confidence intervals is the designing of a record sampling study in which a fixed number of records is reviewed to determine whether a more intensive investigation or an immediate intervention program is necessary. This sort of "lot sampling design" is already a common industrial tool (Miller and Knapp, 1979) that could be modified to account for

repeated samples from the same clinicians (i.e., the same clinician-to-clinician variability issue addressed in the construction of confidence intervals). However, the chance of making an incorrect decision and either abstaining from a needed intervention or implementing an unnecessary one is tied directly to sample size. Hence, any attempt to evaluate individual clinicians requires very large samples. Again, the best recommendation would be to focus on groups of clinicians when global interventions such as education or system improvements can have a powerful impact.

References

Breslow NE, Day NE. Statistical methods in cancer research. Volume 2. The design and analysis of cohort studies. IARC Scientific Publications No. 82. Lyon, France; 1987.

Lawthers A, Palmer RH, Edwards JE, et al. Developing and evaluating performance measures for ambulatory care quality. The Joint Commission Journal on Quality Improvement 1993;19(12):552–65.

Longo DR, Bohr D, editors. Quantitative methods in quality management: a guide for practitioners. Chicago: American Hospital Publishing, Inc.; 1991, p. 136.

Miller MC, Knapp RG. Evaluating quality of care: analytic procedures—monitoring techniques. Rockville, MD: Aspen Publishers Inc.; 1979.

Appendix D. Constructing Algorithm Flowcharts for Performance Measure Evaluation[1]

Writing an Algorithm

An algorithm is a rule of procedure or instructions for solving a problem or accomplishing an objective. Clinical algorithms provide a guide for patient care for a specific problem. They are built on condition-based (branching) logic, in which the condition encountered at a decision node, or point of branching, determines the next pathway. Algorithms in general, and clinical ones in particular, can be represented in words only, and also in a flowchart format.

Clinical algorithms have been published in the English-language medical literature since 1968 (Lusted, 1968). Flowcharts have been shown to be considerably clearer than algorithms in words only for communicating the conditional statements that constitute the underlying logic of most clinical algorithms (Margolis, 1983). They have become the recommended format for representing a clinical algorithm clearly and succinctly.

Guideline and Criteria Algorithms

A clinical practice guideline can be represented as an algorithm written in words only, by expressing its descriptive text in succinct, sequential, declarative statements, as shown in Figure D.1.

When presented in flowchart form, the guideline algorithm becomes a clear description of stepwise recommendations for patient care.

In contrast, review criteria algorithms provide instructions for evaluating criteria conformance for patients in the sample to which the criteria set applies. The criteria assessment algorithm determines whether each of the criteria within the set is met and, if not met, what its status is. The status assigned to each criterion by the algorithm—"met," "not met," "acceptable alternative," "criterion not applicable," "case not reviewable"—indicates whether the care being evaluated adhered to the practice guideline. Criteria algorithms also can be expressed in words only or as flowcharts. Figure D.2 is an example of the

[1]*Author: Naomi J. Banks, M.B.A., M.Ed.*

words-only form of a criterion assessment algorithm based on the same guideline used for Figure D.1. Figure D.3 shows the same algorithm in flowchart form.

Rationale for Flowchart Use

Algorithm flowcharts enhance the usefulness of the algorithm in a number of ways:

- A flowchart provides a visual display of the branching logic of the algorithm.

Figure D.1. Word version of algorithm for a clinical practice guideline (hematocrit example)

Followup of a Low Hematocrit Found by Examining Lab Reports

1. Within 21 days of Hct ≤ 34.0, repeat hematocrit to **confirm anemia.**
2. Within 2 weeks of confirmation of anemia, **work up anemia:**
   —order MCV and reticulocytes
   —or serum iron/TIBC
   —or trial of iron therapy and a repeat hematocrit
3. If iron deficiency anemia is confirmed and the source of blood loss is not known, order three occult blood tests to **search for GI bleeding** within 60 days of initial low hematocrit.
4. If GI bleeding is confirmed, perform rectal exam, sigmoidoscopy, upper GI series, and/or barium enema to **search for GI pathology** within 90 days of the initial low hematocrit.
5. **To follow up on GI workup**, review the results, enter any new diagnosis on the problem list in the chart, and initiate appropriate treatment.

NOTE: GI = gastrointestinal; Hct = hematocrit; MCV = mean corpuscular volume; TIBC = total iron-binding capacity.

Figure D.2. Word version of the algorithm for assessing conformance to review criteria (hematocrit example)

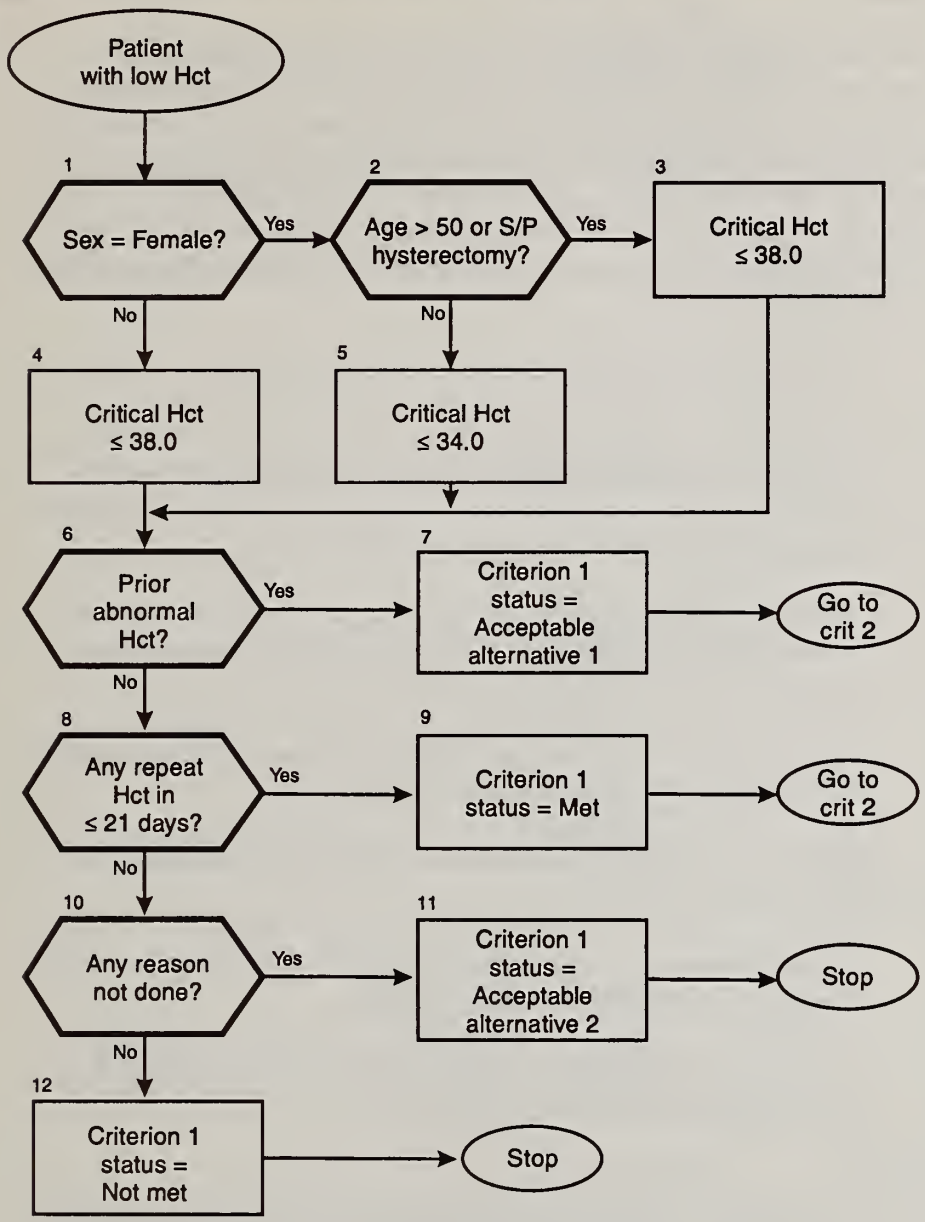
Followup of a Low Hematocrit Found by Examining Lab Reports

Criterion 1: Confirm anemia

1. If the patient is male, a female over 50, or female S/P hysterectomy, the Hct requiring followup is ≤ 38.0, otherwise an index Hct ≤ 34.0.
2. If there was a prior abnormal Hct, the criterion need not be met, the criterion status = "acceptable alternative" 1.
3. If a followup Hct was done within 21 days after an abnormal Hct, the criterion status = "met."
4. If the medical record documents any reason why the criterion could or should not be met, the criterion status = "acceptable alternative" 2.
5. If there was no criterion compliance, the criterion status = "not met."

NOTE: S/P = Status postoperative; Hct = hematocrit.

Figure D.3. Flowchart version of the algorithm for assessing conformance to review criteria (hematocrit example)



NOTE: Hct = hematocrit; S/P = status postoperative

- Working through the logic of the branching at the decision nodes reveals inconsistencies or omissions that may be obscured in an algorithm written in words only.
- All the pathways in the algorithm logic can be traced to determine the appropriateness of the end points.

Flowchart Elements

Shapes

A standard flowchart diagram connects a series of different shapes. Each shape has a specific meaning, as shown in Figure D.4. Although currently there are no universally accepted rules for constructing flowcharts, methods and standards have been proposed (JCAHO, 1992; Leebov and Ersoz, 1991; Margolis, Sokol, Suskind et al., 1992; Pearson, Margolis, Davis et al., 1992). When the shapes of the symbols are consistent within a document, it is much easier to understand the content of the algorithm. The shapes and their names and uses are as follows:

- *Ovals,* the terminal shapes, begin and end the flowchart.
- A *diamond or hexagon,* the decision shape, represents a contingency situation or a question regarding data in the patient record under review. At these decision points, pertinent facts guide the flow to paths representing alternative decisions or actions.
- A *rectangle,* the process shape, represents an action or process and shows what score is assigned to a given criterion based on the path taken.
- *Circles* are the connectors; they indicate where an algorithm continues when it cannot be completed in the space allowed.

Other shapes may take on additional meanings when necessary. Figure D.4 shows these shapes as well:

- A *parallelogram,* which represents a temporary digression to another algorithm from the same practice guideline.
- A *rounded rectangle,* for patient state when clarification of an algorithm pathway is needed.

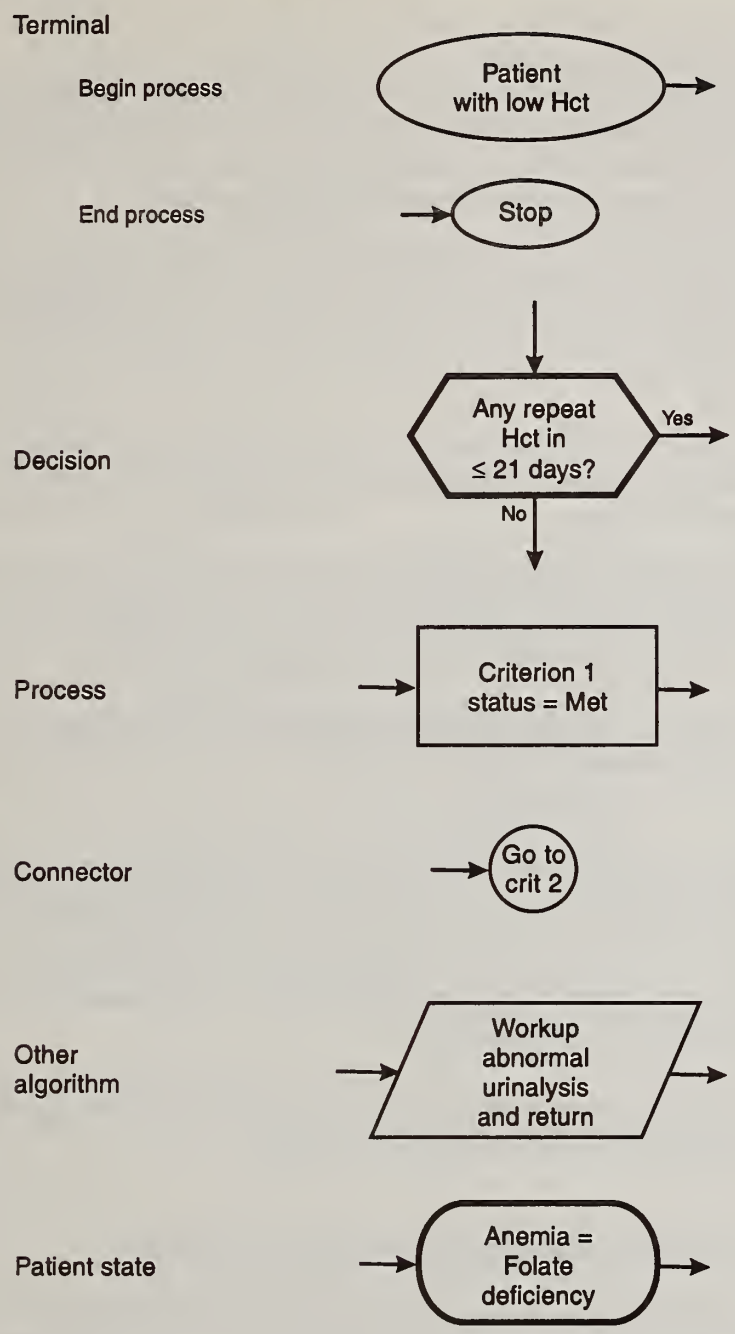
Arrows

Arrows from decision nodes usually flow to the right and down. Traditionally, in an individual flowchart, YES leads in one direction and NO in the other. While the direction of YES and NO may differ among documents, it should always be consistent within a document. When the direction of the arrows is inconsistent, the algorithm logic is difficult to follow.

Numbering

Numbering of shapes on algorithm flowcharts permits the user to identify an action or decision point precisely. In addition, the number may serve as a footnote for reference to notations that define key terms, link care recommenda-

Figure D.4. Common flowchart symbols



tions to published justifications, and consider patient preferences (Hadorn, McCormick, and Diokno, 1992). Numbers should be assigned to shapes sequentially, following horizontal paths first (see Figure D.3, shapes 1 and 2) and then vertical paths (see Figure D.3, shapes 1 and 4). Terminal shapes and connectors are usually unnumbered.

Common Errors in Constructing Flowcharts

The solutions to several types of common flowchart errors are as follows:

- *Lack of consistency within the same flowchart in the direction of the arrows representing YES and NO.* Solution: Reword text in decision nodes or rearrange subsequent decision and action nodes so that all the YES answers lead from the node in the same direction.

- *Using a single flowchart to describe more than one algorithm—for example, one algorithm for both drug treatment decisions and drug inventory control processes.* Solution: Construct separate algorithms, and indicate the entry from one to the other with connector shapes.

- *Giving decision nodes more than two paths outward for multiple choices.* Solution: Construct a series of yes-no decision nodes that cover all options.

- *Giving decision nodes only one exit path, which assumes that there is no other option or no interest in another option.* Solution: Provide alternate path, ending with the terminal shape for "stop."

- *Using the same shape for different meanings or different shapes for the same meaning.* Solution: Review all meanings, and standardize the shapes used for each type.

- *Looping arrows back or exiting to more than one subsequent series of steps, making sequencing hard to follow.* Solution: Use connectors to show a loop or connection to other steps that are drawn separately.

- *Ending a path with a process shape.* Solution: Use a terminal shape to signal the end of each pathway.

- *Writing questions or statements on the connecting arrows.* Solution: Write all text within the appropriate shape.

Choosing Flowcharting Computer Programs

There are several commercially available computer software packages that draw flowcharts. The criteria development process involves numerous drafts of the algorithm flowchart. Using an automated flowcharting program to do this work is considerably more efficient than drawing flowcharts by hand or with a general computer graphics program.

A flowchart program should be selected for its ability to perform basic flowcharting functions, for its ease of editing existing drawings, and for the amount of effort required to learn its basic and advanced features. For projects in which the flowchart will be circulated widely and visual impact is impor-

tant, an additional consideration in software selection is the quality of the printed graphic.

Basic features of flowcharting software programs should permit the user to place a desired shape and write text inside it, connect the shapes with arrows, write titles and move text outside the shapes, construct multiple-page flowcharts, move shapes and text, edit text, and print the flowcharts. Additional features can permit changing the size of shapes, customizing fonts and type size, creating custom shapes, shading shapes and lines, routing lines automatically, changing line thickness, and linking flowchart files with each other or with text files.

References

Hadorn DC, McCormick K, Diokno A. An annotated algorithm approach to clinical guideline development. JAMA 1992;267:3311–4.

Joint Commission on Accreditation of Healthcare Organizations (JCAHO). Using indicator data to improve quality of care, 2d edition. Oakbrook Terrace, IL: JCAHO; 1992.

Leebov W, Ersoz CJ. The health care manager's guide to continuous quality improvement. Chicago: American Hospital Publishing, Inc.; 1991.

Lusted L. An introduction to medical decision making. Springfield, IL: Charles C. Thomas; 1968.

Margolis CZ. Uses of clinical algorithms. JAMA 1983;249(5):627–32.

Margolis CZ, Sokol N, Suskind O, et al. Technical note: proposal for clinical algorithm standards. Med Decis Making 1992;12(2):149–54.

Pearson SD, Margolis CZ, Davis S, et al. The clinical algorithm nosology: a method for comparing algorithm guidelines. Med Decis Making 1992;12(2):123–31.

Glossary

Acceptability. The degree to which health care satisfies patients.

Acceptable alternative. A common and legitimate reason for not conforming to practice guideline recommendations; for example, the clinician recommended a treatment according to the guideline, but the patient refused. Acceptable alternatives are specified explicitly when writing review criteria, whether the alternatives have been stated explicitly in the practice guideline or merely implied. "Acceptable alternative" is also the name of the status assigned to a criterion during a review if documentation is found for a defined acceptable alternative to the criterion.

Accessibility. The ease with which health care can be obtained.

Adverse event. An undesired event implying harm to a patient.

AHCPR. Agency for Health Care Policy and Research, Public Health Service (PHS), Department of Health and Human Services (DHHS).

Algorithm. A rule of procedure, or set of instructions, containing conditional logic for solving a problem or accomplishing a task. Guideline algorithms relate to recommendations for patient care (Gottlieb, Margolis, and Schoenbaum, 1990; Hadorn, McCormick, and Diokno, 1992; Margolis, 1983). Criteria algorithms concern rules for evaluating criteria conformance. Algorithms may be expressed in works only or in diagrammatic form (Margolis, 1992).

AMRRC. American Medical Review Research Center, a private, public interest, and education corporation.

Appropriateness. The probability of benefit to the patients exceeds the probability of harm.

Benchmark. A level of care set as a goal to be attained. Internal benchmarks are derived from similar processes or services within an organization; competitive benchmarks are comparisons with the best external competitors in the field; and generic benchmarks are drawn from the best performance of similar processes in other industries.

Case-based. Refers to a single patient or case.

Case mix. Distribution of a group of patients into categories reflecting differences in patients' diagnoses/conditions.

Case sampling period. The time period during which a case is considered eligible for inclusion in the denominator of a performance measure.

Case severity. A measure of intensity or gravity of a given condition or diagnosis for a patient.

Clinical practice guidelines. Systematically developed statements to assist practitioners' and patients' decisions about health care to be provided for specific clinical circumstances.

Comparative standard. A standard derived from a comparison with other prerformance rates constructed by using exactly the same performance measure, such as the prior performance of a clinician or provider, the observation of the performance of others, or the statistical analysis of group rates.

Confidence interval. An interval or range based on a random sample, for which there is a given probability (e.g., 95 percent) that the population mean is contained within that interval. For example, a study may show that a drug lowers the average blood pressure for patients in the study by 4.8 mm Hg, with the 95-percent confidence interval between 2.5 and 7.3 mm Hg. The confidence interval is used in performance measurement to indicate whether an individual rate from a performance review is considered statistically similar to or different from the group average rate, or from a performance rate selected to represent an acceptable level of care.

Confidence limits. The upper and lower boundaries of a confidence interval.

Criteria (singular: criterion). Standards or principles by which something is judged or evaluated.

Criteria set. A series of criterion statements linked together because they all apply to the same patient sample.

Criterion status. The category to which a case is assigned by application of a criterion. For example, the case that meets a criterion is assigned the status "met"; a case that meets an *acceptable alternative* to a criterion (see definition above) is assigned the status "acceptable alternative." If a criteria set incorporates branching logic (i.e., a certain criterion applies only to a defined subgroup of cases), cases not within that subgroup are assigned the status "not applicable" for that criterion. If data needed to determine whether a case met a criterion are not found in the selected data source, the status "not reviewable" is assigned. A case that does not fit any of the above categories is assigned by default to the status "not met" (i.e., the care given to the case did not conform to the practice guideline).

Data sample. The population or group of patients to whom a performance measure will be applied to assess rates of conformance to a clinical practice guideline. The denominator group.

Decision rule. Instructions for deciding how to evaluate clinical data. When abstracting data from medical records, decision rules determine whether and how a data item should be coded.

DEMPAQ project. The Project to Develop and Evaluate Methods for Promoting Ambulatory Care Quality. A demonstration project with the PRO program of computerized review systems using claims and records data for office-based practice; the computerized review provides comparisons of performance rates to clinician groups to assist them with identifying opportunities for quality improvements.

Denominator. For a performance measure, the sample of cases that will be observed to determine conformance to medical review criteria.

Denominator event/state, index event/state. The event or health state that defines a patient's eligibility for inclusion in the denominator group for a performance measurement.

Efficiency. Occurs when health care of the desired quality is produced at the lowest cost, or when health care produced at a fixed cost is of the highest quality.

Exclusion. Characteristics or conditions that make cases ineligible for review by a specific performance measure or by a specific criterion within a performance measure.

Explicit criteria. Objective criteria specified in advance as a basis for making judgments of performance.

External review. Review in which criteria and standards of judgment are developed or ratified with input from persons other than the clinician or clinician group that is being evaluated.

HCFA. Health Care Financing Administration, Department of Health and Human Services (DHHS).

HCQII. Health Care Quality Improvement Initiative. A program in the pilot testing phase in which peer review organizations (PROs) provide pattern analyses (i.e., comparisons of performance rates) to hospitals to assist them in identifying opportunities for quality improvement.

HSQB. Health Standards and Quality Bureau. The bureau within the HCFA responsible for the Peer Review Organization (PRO) program.

Implicit criteria. Criteria formed by a respected clinician who uses clinical judgment in evaluating performance; these implicit criteria remain concealed in the mind of the reviewer.

Implicit review. Review conducted using implicit criteria.

Index event/state. See **Denominator event/state.**

Indicator. A quantitative measure for monitoring clinical care.

Internal review. Review in which clinicians are involved in setting or adopting the criteria and standards by which they evaluate themselves.

JCAHO. Joint Commission on Accreditation of Healthcare Organizations.

Mean. Arithmetic average of the values of a sample variable.

Measure. An instrument or tool used for making measurements.

Measurement. The performance rate obtained by applying a measure.

Measurement error. Variation in measurements due to causes other than real differences in the attribute being measured.

Medical review criteria. Systematically developed statements that can be used to assess specific health care decisions, services, and outcomes. Each criterion derived from a guideline recommendation is used to determine whether the case being reviewed conforms to a particular recommendation in the guideline. A status is assigned to each criterion to reflect the care given.

Numerator. For a performance measure, the cases in the denominator group that experience events specified in a medical review criterion as evidence of guideline conformance.

Outcome data. Data describing a patient's health status.

Patterns of care. Among a group of clinicians, the distribution of the clinicians' rates of performance.

Peer review. Review conducted by a peer (a similarly qualified clinician) or peers; historically, *peer review* has been conducted by *case-based implicit review,* and so the terms are sometimes used interchangeably.

Performance indicators, indicators. Quantitative measures used to measure and improve performance and quality (JCAHO, in press). *Rate-based* indicators are similar to the *performance measures* defined in Table 2.1; they produce rates for comparing the performance of organizational providers of care. *Sentinel event* indicators identify undesired events such as death; a single instance of a sentinel event triggers a quality review (JCAHO, 1990; p.11; JCAHO, 1991, p. 43). In referring to rate-based indicators, this text uses the term *performance measures,* as does the legislation that established the AHCPR.

Performance measures. Methods or instruments to estimate or monitor the extent to which the actions of a health care practitioner or provider conform to the clinical practice guideline.

Performance rate. A measurement produced by using a performance measure, providing a quantitative evaluation of events related to patient care. A performance rate results when the numerator for a performance measure is divided by the denominator for this measure.

Population-based. The word *population* in this instance is used in the epidemiological sense—a defined group of individuals sampled for study. The term *population-based* is sometimes used to mean *rate-based.* However, a population-based measure generally is a performance rate for a group of patients in a geographic area or in a particular health plan enrollment. When used in this sense, *population-based* applies to all patients rather than to only those who use services.

Practice parameters. Strategies for patient management, developed to assist in clinical decisionmaking.

Prescriptive standard. A statement of what should be achieved rather than a statement of what has been achieved.

PRO. Peer review organization; an organization that reviews appropriateness and quality of care for beneficiaries of the Medicare program.

Process data. What is done to, for, or by patients as part of the delivery of care, such as the performance of a test or procedure.

Profiles. Sets of performance rates aggregated by clinician, clinician group, or organization to monitor some aspect of health care delivery.

Rate. A quantitative measure, usually expressed as a percentage, of the occurence of an event of interest within a specified time interval. Rates are derived by creating a fraction in which the numerator is the number of patients experiencing an event of interest and the denominator is the population of patients at risk for the occurrence of that event. A rate may also be constructed by counting events rather than patients in the numerator and denominator—that is, when the event could occur more than once for a given patient.

Reliability. The extent to which a measurement is reproducible; low levels of random error.

Sample. The subset of a population or the group of cases to whom a performance measure will be applied in order to assess rates of conformance to a clinical practice guideline.

Sensitivity. High rate of detection of "true positives."

Specificity. Low rate of detection of "false positives."

Standard for accreditation. A statement of expectation set by competent authority concerning a degree or level

of requirement, excellence, or attainment in quality or performance (JCAHO, in press).

Standard of care (legal usage). In malpractice case court proceedings there is an attempt to determine whether a patient suffered harm due to negligent violation of a standard of care. The standard of care for the case is elaborated by the questioning of expert witnesses who have studied the facts of the case before the court and have relevant knowledge of comparable behavior.

Standard of care (regulatory usage). Standards for facilities are commonly expressed in terms of a minimal level of policy, equipment, and capacity necessary to achieve licensure or certification.

Standards of quality. Authoritative statements of (1) minimum levels of acceptable performance or results, (2) excellent levels of performance or results, or (3) the range of acceptable performance or results.

Status, criterion. See **Criterion status.**

Structural data. Information about organizational facilities, equipment, policies, and procedures; for example, a hospital policy for patient-controlled analgesia.

Structured implicit review. Implicit review conducted with instructions directing the reviewer to focus on certain types of data or answer certain questions in the review process.

Technical quality. Coordination of judgment, skill, and knowledge in delivering appropriate technology to improve the health of patients.

Threshold. A preestablished level for care. If a desired attribute of care falls below this level or an undesired attribute of care rises above this level, further evaluation or action is triggered.

Timeliness. Services are completed in a timeframe that maximizes health benefit and satisfaction of the patient.

Time window. The time period following an index event during which a case is "observed" for evidence that the care did or did not conform to medical review criteria. In other words, the interval in which it must be determined whether or not a numerator event took place.

Total quality management. A management theory of quality improvement based on (1) involving the total organization, (2) using statistical quality control, (3) seeking to raise the average performance rather than eliminate outliers, and (4) continuously reevaluating performance after interventions in order to plan further interventions if needed.

Unit of analysis. The unit to which a performance measure is applied; the unit may be the patient, clinician, group of clinicians, or institution.

Validity. What is measured, how well it is measured, the effectiveness of a measure in achieving a specific purpose. *Content validity 1* is the degree to which all items in the measure relate to the performance being measured (measure is pure). *Content validity 2* is the extent to which all relevant aspects of performance are covered (measure is complete). *Face validity* is the extent to which a measure appears to measure what it is intended to measure.

Variable. A characteristic that is measured. An *independent variable* is a characteristic that explains another variable. A *dependent variable* is a characteristic that is explained by one or more other variables.

Variation. For performance rates, *variation* refers to differences between the performance rate of one clinician or group of clinicians or organizations and the performance rates of comparable others.

Peer Reviewers[1]

Arja P. Adair, Jr.
Executive Director
Colorado Foundation for Medical Care
Denver, CO

Richard F. Afable, M.D., M.P.H.
Assistant Professor of Medicine
Bowman Gray School of Medicine of Wake Forest University
Winston-Salem, NC

Mary Anne Bacas, R.R.A.
Director of Quality Assessment
Memorial Hospital
York, PA

David Ballard, M.D., Ph.D.
Director
Thomas Jefferson Health Policy Institute
Charlottesville, VA

Daniel M. Barr III, M.D.
Consultant
The MacCullouch Company
Chicago, IL

Chuck Biddle, C.R.N.A., Ph.D.
Assistant Professor of Anesthesiology
Dartmouth–Hitchcock Medical Center
Lebanon, NH

Stephen C. Biondi
Vice President, Quality Assurance and Clinical Services
Unicare Health Facilities
Milwaukee, WI

Catherine Borbas, Ph.D., M.P.H.
Executive Director
Healthcare Education & Research Foundation, Inc.
St. Paul, MN

Troyen A. Brennan, M.D., J.D., M.P.H.
Professor of Law and Public Health
Harvard School of Public Health
Boston, MA

Jerri Bryant, R.N., M.P.H.
Clinical Epidemiologist
Cleveland Clinic Foundation
Cleveland, OH

Richard E. Burney, M.D.
Professor of Surgery
University of Michigan
Ann Arbor, MI

Janet B. Chermak
Vice President of Professional Services
Hillhaven Corporation
Tacoma, WA

Kathleen Ciccone, R.N., M.B.A.
Vice President, Quality Assurance
Hospital Association of New York State
Albany, NY

Kathryn L. Coltin
Director, Clinical Management Information Center
Harvard Community Health Plan
Brookline, MA

Michel Dagher, D.O., M.B.A.
Medical Director, Emergency Department
Memorial Medical Center
Jacksonville, FL

Sidney T. Dana, M.D.
Clinical Professor
SUNY Upstate Medical Center
Syracuse, NY

[1]The individuals listed participated in peer review of a draft version of this document during the summer of 1993. Being listed in this section does not necessarily imply endorsement of this document.

Susan Dean-Baar, Ph.D., R.N.
Assistant Professor
University of Wisconsin–Milwaukee School of Nursing
Milwaukee, WI

Martha Phelps DeMers, R.R.A., M.P.A.
Quality Advisor
Keller Army Community Hospital
West Point, NY

Barbara Demster, M.S., R.R.A.
Vice President
Bottomley & Associates
Gaithersburg, MD

Louis Diamond, M.D.
Georgetown School of Medicine
Washington, DC

Betty T. Dixon, R.N., B.S.N.
Quality Improvement Coordinator
Satilla Regional Medical Center
Waycross, GA

Donalda Ellek
Manager, Office of Quality Assurance
American Dental Association
Chicago, IL

Madelon L. Finkel, Ph.D.
Clinical Professor
Cornell University Medical Center
New York, NY

Kathleen Frawley, J.D., M.S., R.R.A.
Director, Washington, DC Office
American Health Information Management Association
Washington, DC

Peter Goldschmidt
President
World Development Group
Bethesda, MD

Lawrence Gottlieb, M.D., M.P.P.
Associate Medical Director
Harvard Community Health Plan
Brookline, MA

Michael K. Greenberg, M.D.
Member, Quality Standards Subcommittee
American Academy of Neurology
Minneapolis, MN

John F. Griffin, M.D.
Madison–Irving Medical Center
Syracuse, NY

Peter A. Gross, M.D.
Director, Department of Medicine
Hackensack Medical Center
Hackensack, NJ

David Hadorn, M.D.
Consultant
RAND
Santa Monica, CA

Joette Hanna, M.P.A., R.R.A.
Data Manager, Center for Quality Resource Management
Baylor University Medical Center
Dallas, TX

J. Susan Hines
Vice President of Clinical Services and Specialty Programs
Health Care & Retirement Corporation
Toledo, OH

Susie Hutchings, R.N., B.S.N.
Nurse Consultant
Life Care Centers of America
Cleveland, TN

Barbara Irvine, P.A.
Director of Review Services
Oregon Medical Professional Review Organization
Portland, OR

Robert H. Jones, M.D.
Professor of Surgery
Duke University Medical Center
Durham, NC

Douglas G. Kelling, Jr., M.D.
Assistant Professor of Medicine
Duke University
Durham, NC

Jane N. Kimball, R.N.
Director, Nevada Operations
Nevada Peer Review
Las Vegas, NV

Kathy Kunselman, R.N.
Director, Quality Improvement
HealthAmerica of Pittsburgh
Pittsburgh, PA

Judith Lenhart, R.N.
Director, Review Operations
Colorado Foundation for Medical Care
Denver, CO

Steve Levenson, M.D.
Medical Director
Asbury Methodist Village
Gaithersburg, MD

Carole J. Magoffin, M.S.
Executive Director
American Medical Review Research Center
Washington, DC

J. Peter Maselli, M.D.
Medical Director
MassPRO
Waltham, MA

William C. Morgan, M.D.
Medical Advisor to QA/UR Department
Sarasota Memorial Hospital
Sarasota, FL

Alan R. Nelson, M.D.
Executive Vice President
American Society of Internal Medicine
Washington, DC

Karen S. O'Connor, M.A., R.N.
Director of Practice, Economics, and Policy
American Nurses Association
Washington, DC

Denis M. O'Day, M.D.
Chair, Department of Ophthalmology and Visual Sciences
Vanderbilt University School of Medicine
Nashville, TN

Delwin K. Ohrt, M.D.
Vice President/Medical Director
Blue Cross/Blue Shield of Minnesota
St. Paul, MN

Patricia A. Pickering, M.S.M., R.N., C.P.N.P.
Quality Management Specialist
Healthnet, Inc.
Indianapolis, IN

Julia A. Powell, B.S.N., M.A., C.N.A.
Vice President, Patient Services
National HealthCorp. L.P.
Murfreesboro, TN

Richard G. Roberts, M.D., J.D.
Associate Professor, Department of Family Medicine
University of Wisconsin
Madison, Wisconsin

Henry D. Royal, M.D.
Chairman, Office of Health Care Policy
Society of Nuclear Medicine
Elmwood Park, NJ

Robert Sebring, Ph.D.
Director, Division of Quality Care
American Academy of Pediatrics
Elk Grove Village, IL

Earl E. Smith III, M.D.
American College of Emergency Physicians
Dallas, Texas

L. Kent Smith, M.D., M.P.H.
Medical Director, Preventive Medicine Program
Arizona Heart Institute
Phoenix, AZ

Carl E. Speicher, M.D.
Professor and Director, Clinical Services
Ohio State University Hospitals
Columbus, OH

Marilyn P. Verhey, Ph.D., R.N.
Assistant Professor
San Francisco State University
San Francisco, CA

Jonathan Warren, M.D.
Chairman, Guidelines Committee
American College of Critical Care Medicine
Anaheim, CA

Rhonda Whitson, R.R.A.
Project Administrator, Practice Management Department
American College of Emergency Physicians
Dallas, TX

Billie Ann Young, M.B.A., R.R.A.
Chairman-Elect, Quality Assurance Section
American Health Information Management Association
Boca Raton, FL

Wanda W. Young, Sc.D.
President
The Pittsburgh Research Institute
Pittsburgh, PA

Les Zendle, M.D.
Associate Medical Director
Kaiser Permanente
Pasadena, CA

*U.S. G.P.O.:1995-380-940

U.S. Department of Health and Human Services
Public Health Service
Agency for Health Care Policy and Research
Executive Office Center, Suite 501
2101 East Jefferson Street
Rockville, MD 20852

AHCPR Publication No. 95-0046
March 1995